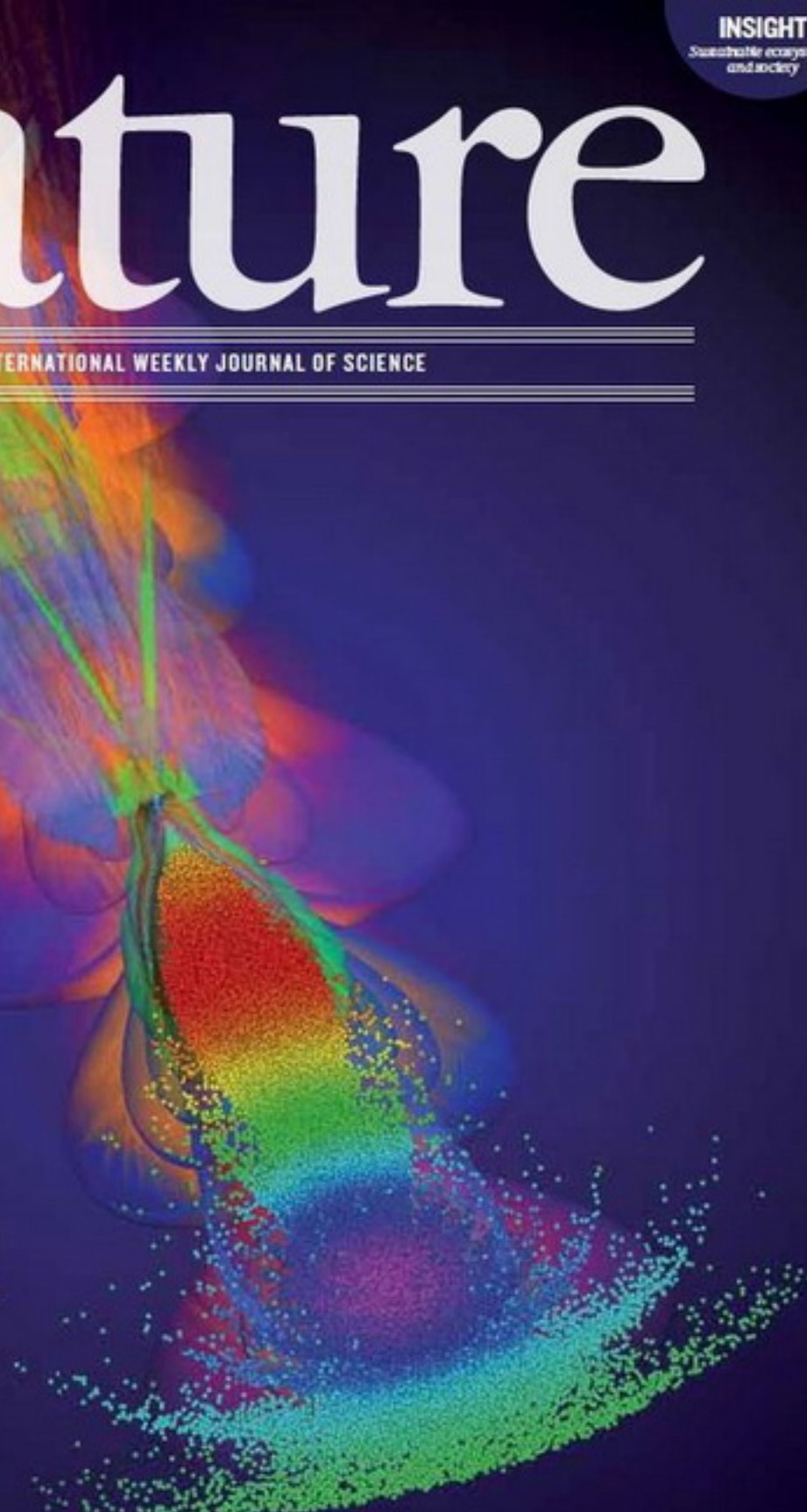


# nature

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE

## FULL SPEED AHEAD

Plasma wakefield  
machines — the particle  
accelerators of the  
future? **PAGES 40 & 92**



### CENTRAL EUROPE

#### LIFE AFTER THE WALL

Science 25 years after the  
collapse of communism

**PAGE 22**

### ENVIRONMENTAL SCIENCE

#### CASH, CONFLICT, CONSERVATION

To protect the planet, stop the  
infighting and fudging

**PAGES 27, 28 & 32**

### INTERACTIVE NOTEBOOKS

#### SHARE AND SHARE ALIKE

IPython allows  
on-the-run analysis

**PAGE 151**

**NATURE.COM/NATURE**

6 November 2014 £10

Vol. 515, No. 7525



9 770028 083095



# THIS WEEK

## EDITORIALS

**CONSERVATION** Saving species is far from a walk in the park **p.8**

**WORLD VIEW** Psychology gears up to check its workings **p.9**



**BREAKFAST** Chimps plan days to ensure they nab tastiest figs **p.11**

## Journals unite for reproducibility

*Consensus on reporting principles aims to improve quality control in biomedical research and encourage public trust in science.*

**R**eproducibility, rigour, transparency and independent verification are cornerstones of the scientific method. Of course, just because a result is reproducible does not make it right, and just because it is not reproducible does not make it wrong. A transparent and rigorous approach, however, will almost always shine a light on issues of reproducibility. This light ensures that science moves forward, through independent verifications as well as the course corrections that come from refutations and the objective examination of the resulting data.

It was with the goal of strengthening such approaches in the biomedical sciences that a group of editors representing more than 30 major journals; representatives from funding agencies; and scientific leaders assembled at the American Association for the Advancement of Science's headquarters in June 2014 to discuss principles and guidelines for preclinical biomedical research. The gathering was convened by the US National Institutes of Health, *Nature* and *Science* (see *Science* **346**, 679; 2014).

The discussion ranged from what journals were already doing to address reproducibility — and the effectiveness of those measures — to the magnitude of the problem and the cost of solutions. The attendees agreed on a common set of Principles and Guidelines in Reporting Preclinical Research (see [go.nature.com/ezjl1p](http://go.nature.com/ezjl1p)) that list proposed journal policies and author reporting requirements in order to promote transparency and reproducibility.

The guidelines recommend that journals include in their information for authors their policies for statistical analysis and how they review the statistical accuracy of work under consideration. Any imposed page limits should not discourage reproducibility. The guidelines encourage using a checklist to ensure reporting of important experimental parameters, such as standards used, number and type of replicates, statistics, method of randomization, whether experiments were blinded, how

the sample size was determined and what criteria were used to include or exclude any data. Journals should recommend deposition of data in public repositories, where available, and link data bidirectionally when the paper is published. Journals should strongly encourage, as appropriate, that all materials used in the experiment be shared with those who wish to replicate the experiment. Once a journal publishes a paper, it assumes the obligation to consider publication of a refutation of that paper, subject to its usual standards of quality.

***“The guidelines encourage using a checklist to ensure reporting of important experimental parameters.”***

The more open-ended portion of the guidelines suggests that journals establish best practices for dealing with image-based data (for example, screening for manipulation, storing full-resolution archival versions) and for describing experiments in full. An example for animal experiments is to report the source, species, strain, sex, age, husbandry

and inbred and strain characteristics for transgenic animals. For cell lines, one might report the source, authentication and mycoplasma contamination status. The existence of these guidelines does not obviate the need for replication or independent verification of research results, but should make it easier to perform such replication.

Some of the journals at the meeting had already had all or most of these principles and guidelines in place. But the point is that a large number of scientific journals are standing together in their conviction that reproducibility and transparency are important issues. As partners to the research enterprise in the communication and dissemination of research results, we want to do our part to raise the standards for the benefit of scientists and of society. The hope is that these guidelines will be viewed not as onerous, but as part of the quality control that justifies the public trust in science. ■

## On the mend

*The scientific regeneration of central Europe is gathering pace, but needs further help to thrive.*

**T**he peaceful implosion of communism in the autumn of 1989, almost exactly 50 years after Nazi Germany's assault on Poland triggered the Second World War, was perhaps the brightest moment in Europe's twentieth-century history. The fall of the Berlin Wall restored political and personal freedom to central Europe, where people had endured Hitler's atrocities only to find themselves ruled by Soviet despots. It is a small miracle that the rich learned tradition of the region survived two consecutive tyrannies.

Science in liberated central Europe had to adapt quickly to survive in the free world. Governments, intellectual elites and academic institutions in the region were all equally unprepared for the political sea change that occurred after 1989. A quarter of a century on, the transformation to parliamentary democracy and a market economy has been achieved. Science was generally not a priority in the early years of the transition. But from 2004 onwards, membership of the European Union (EU) provided a boon that some countries are prudently using to rebuild their research capacities (see page 22). However, despite generous subsidies from Brussels, other countries have a long way to go.

The region's main asset is a growing pool of young talent that is rediscovering science as a worthwhile profession. This generation rightly demands more support and more constructive political vision than some of the region's current governments have to offer. In Romania and Bulgaria, where those in power are stubbornly obstructing reform, science is losing out. And in Hungary,



where pluralism is under threat, the writing is on the wall.

The last generation of scientists who trained under the communist regime had little faith in science in their countries. When freedom arrived, most of them grasped vastly more lucrative business opportunities or chose to migrate permanently to the West. The resulting shortage of mid-career scientists can be seen in almost any research department in the region. It is also the main reason why central Europe attracts shamefully few grants from the European Research Council and why the region is lagging behind in terms of its overall scientific output (see graphic on page 24).

But interest in science and higher learning is rising sharply. In Poland, student numbers have increased fivefold since 1990. Overall, more than one-quarter of the 20 million students in the EU are now from the new member states. The vast majority of them were born after the demise of communism.

Young scientists do continue to leave, and so they should, but the excessive brain drain has thankfully ceased. Most countries in the region have in recent years created transparent and strictly merit-based science funding systems. But a lot more needs to be done — especially at universities — to help talented young scientists gain independence at an early age. Institutes that do employ young independent group leaders, such as the International Institute of Molecular and Cell Biology in Warsaw, are reaping the benefits. Furthermore, funding agencies across the region should expand, and better advertise, programmes aimed at repatriating young foreign-trained scientists.

Universities in the region are a long way from scoring in the upper ranks of international academic comparisons. But the idea that central Europe remains a poor relation in global science is obsolete.

Since 2004, numerous labs have been re-equipped to facilitate competitive science. The best institutes now offer conditions on a par with those at aspiring labs in Singapore, China or Saudi Arabia. But science managers in central Europe lack the bravado with which the new players in Asia and the Middle East trumpet their strengths and the aggressiveness with which they recruit foreign talent.

The new European Commission that takes office this month must help the region to raise its scientific profile. The EU's €80-billion (US\$100 billion) Horizon 2020 programme, which started this year, includes a scheme that invites less-potent member states to open new research centres, or upgrade existing ones, in partnership with richer countries. Leading research institutions in the West should accept the invitation.

Collaborations involving high-profile British universities, say, or the prestigious German Max Planck institutes, would no doubt raise the visibility of central European science and help to improve the region's participation in EU-funded research. Structural funds will also remain essential. Billions have already been earmarked for the 2014–20 period, and research is to remain a major beneficiary. But the commission should closely monitor the effectiveness of the investment.

Science is at the heart of the EU's policies. Its renaissance in many parts of central Europe serves as an example of successful European integration at a time when forces threatening the EU's social and political cohesion are gaining strength. Domestic neglect of science in the continent's southeast risks casting the EU's poorer countries even further adrift from the rest of Europe. ■

# Protect the parks

*Balancing the needs of development and conservation is difficult — but urgent.*

The World Parks Congress could not be in a better location this year than Sydney, Australia. The gathering of researchers, policy experts and conservationists occurs only once a decade, and it arrives in Australia next week as debate over protection of one of the most famous parks in the nation — and the world — reaches fever pitch.

The Great Barrier Reef is celebrated as one of nature's true wonders. In the popular imagination it is the quintessential coral reef. Millions visit every year to swim in its waters, and millions more dream of such a trip. It is rightly a source of pride for many Australians.

But the reef is in trouble, beset by direct human activity in the form of coastal development and indirect activity in the form of climate change. It exemplifies the problems that congress attendees must struggle with.

The International Union for Conservation of Nature — which oversees the meeting — says that the 2003 parks congress in Durban, South Africa, led to major advances, including ongoing work on protected areas under the Convention on Biological Diversity, and revisions to the system used to manage protected areas.

The world has changed since 2003, when political rhetoric around climate change was still gearing up. Levels of carbon dioxide in the atmosphere — measured at the US National Oceanic and Atmospheric Administration's Mauna Loa observatory in Hawaii — have increased from an average of 375.77 parts per million (p.p.m.) to current peaks of more than 400 p.p.m. on some days, the highest for thousands of years.

It is true that there has been notable progress in creating protected areas, especially in the oceans: huge parks were established in the Pacific by US President George W. Bush and then expanded by President Barack Obama. These join similar areas, including one around the Chagos Islands in the Indian Ocean, created by the United Kingdom.

Progress towards internationally agreed targets to protect 10% of the world's oceans is slow and behind schedule, but it is progress nonetheless. And growing environmental movements are making themselves heard in rapidly developing economies such as China. In this issue of *Nature*, seven researchers set out their vision of how the world's parks should develop (page 28). This work is more important than ever.

There have been setbacks. Extensive networks of parks across Africa have failed to safeguard rhinoceroses and elephants from surging poaching. Deforestation continues apace. And the Great Barrier Reef itself has been hard hit, with coral cover falling at a worrying rate (see page 16).

Ecosystems are complex and ever-changing, and it is difficult to prove that protections make a difference. Well-funded and well-organized business lobby groups seek to develop many important areas, and development often still trumps wildlife. People quite legitimately want to improve their own lot, even at the expense of other species. Wealthy nations are able to seal off certain areas for protection, but developing nations in which people rely on subsistence hunting and fishing may not have that luxury.

How to measure the value of the world's ecosystems is still debated. Edward Barbier argues on page 32 that economists tracking growth have done us a disservice by omitting from their calculations the value that has been lost by, for example, turning a mangrove forest into a shrimp farm.

And conservationists still argue over why we should protect the planet — for its economic value or for its own sake. On page 27, Heather Tallis and Jane Lubchenco say that “vitriolic” battles over this tension endanger conservation science. “It is time to re-focus the field of conservation on advancing and sharing knowledge in all relevant disciplines and contexts, and testing hypotheses based on observations, experiments and models,” they write in a petition with 238 other signatories.

In conservation, it is often not clear when deadlines are until they have passed. At the next World Parks Congress, around 2024, what will attendees discuss? Will there be any truly wild rhinos left? Will the Great Barrier Reef be in terminal decline? Hopefully not. But this year's attendees have their work cut out. ■

➔ **NATURE.COM**  
To comment online,  
click on Editorials at:  
[go.nature.com/xhunqy](http://go.nature.com/xhunqy)





## Metascience could rescue the 'replication crisis'

*Independent replication of studies before publication may reveal sources of unreliable results, says Jonathan W. Schooler.*

If you witness a bank robbery, what should you do? Tell the police what you saw? And then volunteer to identify the arrested suspect in a line-up? That sounds like the right response, but it might not be. Almost 25 years ago, a colleague and I published a psychology study (J. W. Schooler and T. Y. Engstler-Schooler *Cogn. Psychol.* 22, 36–71; 1990) that indicated that to describe the physical appearance of the criminal could make it harder for a witness to subsequently identify the person they saw.

The effect is called 'verbal overshadowing' and its discovery proved controversial, and not simply because of what it means for detectives. Over the years, other researchers (myself included) have had mixed success replicating the finding. Some have doubted whether it exists at all. That 1990 research project has been used as an example of the 'replication crisis' that has engulfed science in recent years. In disciplines such as medicine, psychology, genetics and biology, researchers have been confronted with results that are not as robust as they originally seemed.

In response, scientists have launched various replication projects to assess the robustness of published research. In psychology, numerous labs have volunteered to re-run studies, with the methods vetted by the original researchers. My discovery of the verbal-overshadowing effect was an obvious target for this approach, and so, last year, psychologists at 31 different laboratories across the world signed up to repeat the study and report the results.

How did I feel about having work scrutinized in this way? I thought that I could not lose. Positive replication would confirm the important verbal-overshadowing effect. Failure to replicate would be more evidence for the 'decline effect', an idea I endorse that the size of an effect decreases over repeated replications, for reasons that are not fully understood.

Unfortunately for a replication study, there was a mistake in the timing parameters of the initial experimental protocol. Still, this unexpected negative turn of events took a positive spin as the deviation from the original protocol, once identified and fixed, generated some useful comparative data.

Some 22 of the original 31 repeating labs went on to follow the corrected study protocol. Pooled, the results confirmed the original finding. The verbal-overshadowing effect was clearly demonstrated (although the effect size was smaller than in our 1990 research and observed only when the original parameters, in the corrected protocol, were followed).

The outcome is a genuine victory for the emerging field of metascience, an approach in which science turns the lens of scrutiny on itself. Metascience, the science of science, uses rigorous methods to examine how scientific practices

influence the validity of scientific conclusions. It has its roots in the philosophy of science and the study of scientific methods, but is distinguished from the former by a reliance on quantitative analysis, and from the latter by a broad focus on the general factors that contribute to the limitations and successes of research.

Large-scale replication efforts such as this one are important, but they are expensive, time-consuming and impracticable for the vast majority of scientific studies. Rather than focus exclusively on whether past studies stand up, we need a clearer sense of the processes that influence the reliability of new findings. These could include how 'invested' researchers are in the original hypothesis, the number of times a protocol is repeated and how the methods and outcomes are assessed and written up.

Together with labs at three other universities, my research group has begun an initiative to test the reproducibility of our science. Rather than re-examine published studies, each lab has agreed to allow the other three to generate new findings, replicate each others' experiments and compare the results of new studies before they are published. We will then be able to judge differences, for example, in the results obtained by the originating lab compared with those that follow the idea, and so, in theory, have less invested in the results.

It is clearly not feasible for all researchers to follow this approach in their routine work. But it should offer a valuable academic exercise to examine the factors that affect reproducibility as they arise during the course of research.

By pre-registering all aspects of new scientific studies and then repeatedly trying to replicate them, the project allows careful scrutiny of all parts of the research process, from inception to replication. If the studies replicate flawlessly, we will have established a gold standard for reproducible studies. If they do not, then our approach will present an opportunity to rigorously assess the reasons.

Some might suggest that the focus on replication within psychology is an indictment of the field. It is precisely the opposite. All fields face problems with reproducibility, and psychology should be applauded for its willingness to tackle the issue empirically.

In fact, psychological science has long been at the forefront of refining and improving the scientific process. The understanding of experimenter expectancy effects (a form of cognitive bias) and the importance of double-blind trials emerged first in psychology. Such self-examination can only strengthen the scientific process for all. ■

**Jonathan W. Schooler** is a psychologist at the University of California, Santa Barbara.  
e-mail: [jonathanwschooler@gmail.com](mailto:jonathanwschooler@gmail.com)

**WE NEED A  
CLEARER  
SENSE OF THE  
PROCESSES  
THAT INFLUENCE THE  
RELIABILITY  
OF NEW  
FINDINGS.**

➔ **NATURE.COM**  
Discuss this article  
online at:  
[go.nature.com/8iabws](http://go.nature.com/8iabws)



## CLIMATE CHANGE

## Greenland's ice at mercy of ocean

As the climate warms, changes to the ocean around Greenland could cause its ice sheet to melt where it currently seems stable.

Camille Lique and her colleagues at the University of Oxford, UK, used data from an integrated climate model to study a worst-case scenario: concentrations of atmospheric carbon dioxide that increase by 2% each year for 70 years until they are four times the current level. The authors found that the resulting changes in ocean circulation will contribute to ocean warming across the entire region, by up to 5°C in places. This will lead to increased melting of marine-terminating glaciers across most of Greenland.

The potential for widespread ice loss suggests that ice-sheet monitoring should not be limited to areas of current, rapid melting, the authors say. *Clim. Dynam.* <http://doi.org/ws4> (2014)

## ANIMAL BEHAVIOUR

## Chimps plan for better breakfasts

Wild chimpanzees plan their days to improve the chance of finding tasty fruit for breakfast.

Chimps (*Pan troglodytes verus*) like ripened figs (pictured), but these treats are available only for short periods of time and are sought by other animals. To find out how chimps secure the prized fruit, Karline Janmaat and her colleagues at the Max Planck Institute for Evolutionary Anthropology in Leipzig,

Germany, monitored five wild female chimps at Tai National Park in the Ivory Coast for 275 days over 2 years. They found that when figs were ripe, the animals often left their bed nests before dawn, and departed earlier when the fig tree was farther away.

Such flexible planning may have supported the evolution of calorie-hungry big brains in other primates and ancient human ancestors, the researchers say.

*Proc. Natl Acad. Sci. USA* <http://doi.org/ws6> (2014)

## NEUROSCIENCE

## Nostalgia rewards the brain

Reminiscing about happy times is rewarding to the brain, and people will even give up money for the chance to enjoy some nostalgia.

Mauricio Delgado and his colleagues at Rutgers University in Newark, New Jersey, asked volunteers to recall happy and neutral memories while their brains were scanned using functional magnetic resonance imaging. Participants spent more time recalling happy memories, and when doing so, their brain activity patterns were similar to those seen in people receiving money. When offered a small amount of money to recall a positive memory and a larger amount for a neutral memory, the volunteers were more likely to choose the happy memory.

The researchers say that recalling good memories could be useful for improving mood. *Neuron* <http://doi.org/ws2> (2014)

## ATMOSPHERIC SCIENCE

## Warming from soot overestimated

Atmospheric soot may not have nearly as much climate warming potential as previously thought.

Tiny carbon particles produced by biomass burning and incomplete combustion of fossil fuels absorb sunlight, warming the planet. Xuan

## SOCIAL SELECTION

Popular articles on social media

## Conference gender gap revealed

Conferences are a central part of scientific life, but they are also an arena for gender disparities, according to a study proving popular on social media. Researchers in Australia gathered data from the 2013 Australasian Evolution Society meeting and found that male speakers tend to get a bigger share of the exposure — a conclusion shared by past studies of conferences. Even though roughly the same number of men and women attended and presented at the evolution conference, women spoke for less time and were also less inclined to ask for longer slots for their talks. Katie Hinde, an evolutionary biologist at Harvard University in Cambridge, Massachusetts, shared her take-home message on Twitter: “Ladies, request the long talk.” *PeerJ* 2, e627 (2014)



Based on data from altmetric.com. Altmetric is supported by Macmillan Science and Education, which owns Nature Publishing Group.

➔ **NATURE.COM**  
For more on popular papers:  
[go.nature.com/qcd1tf](http://go.nature.com/qcd1tf)

Wang at the Massachusetts Institute of Technology in Cambridge and his colleagues used a refined chemical transport model, along with actual observations, to study black-carbon behaviour.

They found that previous simulations have substantially overestimated concentrations of soot in remote regions as well as its global lifetime, leading to higher estimates of its warming potential. The authors conclude that the direct warming effect of black carbon might be less than one-quarter of the previously reported value.

Policies aimed at reducing black-carbon emissions could have only a limited impact on mitigating climate warming, the team cautions.

*Atmos. Chem. Phys.* 14, 10989–11010 (2014)

form of hydrogen.

The molecule BrHBr is held together by weak electrostatic attractions known as Van der Waals' forces. Jörn Manz at Shanxi University in Taiyuan, China, and his colleagues calculated what would happen if the hydrogen were swapped for a lighter isotope called muonium, in which a positively charged elementary particle called an antimuon takes the place of the proton.

They predict that the BrMuBr molecule would be held together not by electrostatic forces but with a vibrational bond. The muonium shuttling between the bromine atoms would form a lower-energy system than the vibrations of MuBr alone.

These calculations suggest that the bond might have been produced in the earlier experiment, which combined muonium and bromine. *Angew. Chem. Int. Ed.* <http://doi.org/f2vjn6> (2014)

➔ **NATURE.COM**  
For the latest research published by Nature visit:  
[www.nature.com/latestresearch](http://www.nature.com/latestresearch)

## Vibrations yield new type of bond

Calculations suggest that a new kind of chemical bond proposed in the 1980s might have occurred in a 2012 experiment that coupled two bromine atoms to an exotic



FERRERO-LABAT/ARDEA.COM



# SEVEN DAYS

The news in brief

## POLICY

### One of the gang

Poland signed an agreement on 28 October to join the European Southern Observatory (ESO). The country is set to become the ESO's 14th European member state once its parliament ratifies the agreement. Membership of the organization will give Polish astronomers access to the ESO's ground-based telescopes. It will also allow companies in Poland to bid for contracts to build the 39-metre European Extremely Large Telescope, which is being constructed in Cerro Armazones in Chile. Brazil, the ESO's only non-European member, has yet to ratify an accession agreement that it signed in 2010.

### Climate warning

The Intergovernmental Panel on Climate Change (IPCC) has warned of "severe, pervasive and irreversible impacts for people and ecosystems" if greenhouse-gas emissions are not substantially reduced over the next few decades. The warning, addressed to policy-makers, is included in the summary of the IPCC's fifth assessment report of climate risks, released on 2 November in Copenhagen. The summary distills the latest contributions by the IPCC's three working groups to the fifth assessment, as well as two special reports.

### Vaccine approval

US regulators on 29 October approved a vaccine against a deadly strain of a bacterium that causes meningitis. Trumenba, produced by a subsidiary of Pfizer of New York, fends off infection by *Neisseria meningitidis* serogroup B — a class of bacterium known for

causing outbreaks among university students. Last December, the Food and Drug Administration allowed the use of a different meningitis vaccine that is not approved in the United States — but is available in Europe — at universities where outbreaks had occurred.

### Polar talks fail

The Commission for the Conservation of Antarctic Marine Living Resources concluded its annual meeting in Hobart, Australia, on 31 October, without agreeing on a plan to create a massive marine reserve in the Ross Sea. The commission, made up of representatives from 24 countries and the European

Union, has failed 3 times before to agree on similar plans to ban fishing in what some researchers say is the most endangered area of the polar region (see *Nature* <http://doi.org/wtt; 2014>).

### Emissions fines

The US Environmental Protection Agency announced on 3 November that car makers Hyundai and Kia will pay the largest settlement ever for alleged violations of the US Clean Air Act. The companies are said to have sold nearly 1.2 million vehicles that will collectively emit about 4.75 million tonnes of greenhouse gases above the amount that the companies certified to the agency. In

addition to a US\$100-million penalty, they will also forfeit 4.75 million previously claimed greenhouse-gas emissions credits, estimated to be worth more than \$200 million.

## EVENTS

### Spaceflight crashes

US government investigators are probing the 31 October crash of Virgin Galactic rocketplane *SpaceShipTwo*, which killed one of the craft's two pilots. The plane was on a test flight for commercial space travel. Three days earlier, a crewless Antares rocket exploded seconds after being launched from Wallops Island, Virginia, destroying scientific equipment and experiments



USGS/GETTY

## Lava invades Hawaiian town

A creeping lava flow from the volcano Kilauea on Hawaii is threatening a community of almost 1,000 people. The flow (pictured), which began on 27 June, has travelled roughly 20 kilometres and reached the town of Pahoa, where it has overrun pastures, a cemetery and private property on a course that is heading for the town's main road. The US National Guard was

deployed last week to help to erect a roadblock, and about 20 families were told to evacuate their homes. By 30 October, the leading edge had stalled 155 metres from the road, but on 2 November, the US Geological Survey reported active lava breakouts from parts of the flow. Kilauea is the most active volcano on Hawaii; its current eruption began in 1983.

headed for the International Space Station. Orbital Sciences in Dulles, Virginia, which operated the Antares, is one of only two private companies with a NASA contract to fly cargo to the space station. See page 15 and *Nature* <http://doi.org/ws5> (2014) for more.

## RESEARCH

## Political research

On 28 October, the presidents of Stanford University in California and Dartmouth College in Hanover, New Hampshire, issued a public apology for a controversial political-science experiment in Montana. The study aimed to test whether voters' decisions would be affected by pre-election flyers that characterized the political attitudes of state Supreme Court candidates. Roughly 100,000 people received the flyer, which appeared to be an official government document because it bore the state's seal. Stanford said that the research had not been submitted for approval by its institutional review board; both universities are investigating other possible violations.

## Lunar loop

China successfully completed its first robotic mission to the Moon and back on 1 November. Launched on 23 October, the probe



flew around the Moon and survived re-entry into Earth's atmosphere to land safely in Inner Mongolia (**pictured**). The vehicle, nicknamed Xiaofei, or little flyer, had no scientific goals, but was intended to test technology for Chang'e-5, a mission to return lunar samples to Earth planned for 2017. The success makes China the first country to fly a probe around the Moon and back since the Soviet Union in the 1970s.

## Falsified data

The US Office of Research Integrity on 29 October reported findings that a former laboratory director at the US National Institute of Arthritis and Musculoskeletal and Skin Diseases in Bethesda, Maryland, had committed research misconduct. An investigation concluded that Bijan Ahvazi, previously the head of the Laboratory

of X-ray Crystallography, had falsified data related to three publications. As part of a settlement, Ahvazi has agreed to have his research supervised and to be excluded from peer-review committees for agencies such as the US National Institutes of Health for two years.

## FUNDING

## Funding boon

Germany's universities and large science organizations have been promised a boost to help them to cope with rising student enrolment and the increasing costs of research. At their meeting on 30 October, science ministers from the federal government and Germany's 16 state governments pledged €25.3 billion (US\$31.6 billion) over the next six years to continue special programmes for science and higher

## COMING UP

### 8–9 NOVEMBER

On the 25th anniversary of the fall of the Berlin Wall, the city holds the sixth Falling Walls Conference. Scientists from around the world discuss impending breakthroughs in areas including cancer care, neuroengineering and global energy.

[go.nature.com/beyrwm](http://go.nature.com/beyrwm)

### 12 NOVEMBER

The European Space Agency's Rosetta craft attempts to land a washing-machine-sized probe on the surface of comet 67P/C-G.

### 12–19 NOVEMBER

The World Parks Congress takes place in Sydney, Australia, featuring sessions on the use of mobile technology in conservation, successes and challenges in rhino conservation and the future of privately protected areas.

[go.nature.com/c51ghb](http://go.nature.com/c51ghb)

education. See [go.nature.com/6qsa5n](http://go.nature.com/6qsa5n) and page 17 for more.

## Money for malaria

The Bill & Melinda Gates Foundation announced on 2 November a US\$156-million contribution to the PATH Malaria Vaccine Initiative. The donation will support two lines of development: vaccines that prevent people from becoming infected after being bitten by infected mosquitoes, and transmission-blocking vaccines that prevent mosquitoes from becoming infected when they bite people with malaria — with special attention to vaccines that combine both features.

**NATURE.COM**

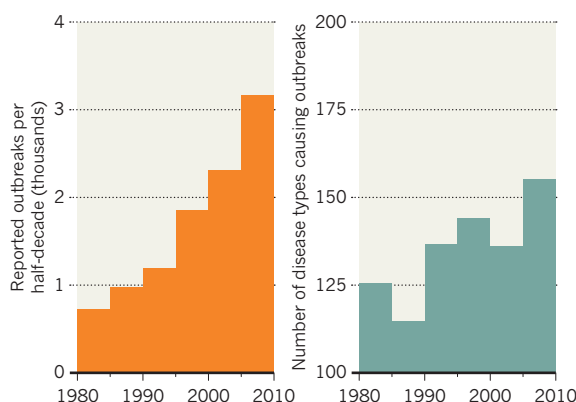
For daily news updates see:  
[www.nature.com/news](http://www.nature.com/news)

## TREND WATCH

A global analysis suggests that human infectious-disease outbreaks are becoming more frequent and more diverse (K. Smith *et al.* *J. R. Soc. Interface* **11**, 20140950; 2014). A team led by Katherine Smith of Brown University in Providence, Rhode Island, found that the trends were significant even after correcting for changes in surveillance and reporting (for example, using indirect measures such as Internet use). However, the number of cases per person is falling, they find.

## DISEASE OUTBREAKS ON THE RISE

Both the number and diversity of outbreaks of human infectious diseases have risen since 1980.





# NEWS IN FOCUS

**CONSERVATION** Future of Great Barrier Reef splits scientists **p.16**

**EBOLA** Models cannot explain apparent plateau in Liberian cases **p.18**

**RELIGION & SCIENCE** Fossil-hunting creationist sues university **p.20**

**CENTRAL EUROPE** The state of science 25 years since the Berlin Wall fell **p.22**



RINGO H. W. CHIU/AP



Virgin Galactic's *SpaceShipTwo* rocket plane crashed on 31 October, scattering wreckage across the Mojave Desert in California.

## COMMERCIAL SPACEFLIGHT

# Fledgling space industry resolute after fatal crash

*Virgin Galactic accident should not be allowed to stifle innovation, warn analysts.*

BY ALEXANDRA WITZE

What was already a bad week for commercial spaceflight turned tragic on 31 October with the fatal crash of *SpaceShipTwo*, the rocket plane owned by Virgin Galactic that was meant to ferry tourists and scientists to the edge of space. The fledgling private space industry is reeling from the accident, in which the co-pilot, Michael Alsbury, died, and the pilot, Peter Siebold, was left seriously injured. The crash came just three days after the explosion of an uncrewed Antares rocket that was launching on a commercial mission to resupply the International Space Station.

But analysts warn against letting the accidents

curb spaceflight innovation. "It's absolutely critical that people don't throw up their hands and say 'it's just too hard,'" says Joan Johnson-Freese, a space-policy specialist at the US Naval War College in Newport, Rhode Island.

As *Nature* went to press, the cause of the accident was still unclear. A 2 November press conference held by the US National Transportation Safety Board (NTSB) focused on the tail booms. These are intended to help slow the rocket plane's descent when it re-enters the atmosphere, in a process known as feathering. Pilots are supposed to engage the feathering action when the craft is moving at roughly 1.4 times the speed of sound, at the top of its parabolic arc of flight. The idea, inspired by a

badminton shuttlecock, is to increase drag and allow the spaceship to descend safely.

Two steps are normally required for feathering, said Christopher Hart, acting chairman of the NTSB, which is holding its first accident investigation involving the commercial space industry. On the fateful *SpaceShipTwo* flight, video footage retrieved from the cockpit shows that one of the pilots completed the first step when the vehicle was moving at only the speed of sound. The second action was not performed, but seconds later, the tail booms began moving to their feathered position anyway.

Until that point, *SpaceShipTwo* had performed as expected. It took off from the Mojave Air and Space Port in California, ►

► hanging beneath its carrier aeroplane. It then detached, and fired its rocket motor. Within minutes, it had disintegrated and its pieces were scattered over an 8-kilometre-long swathe of the Mojave Desert.

Among other goals, the flight was meant to test a new type of rocket motor. Virgin Galactic decided in May to switch *SpaceShipTwo*'s fuel from a rubber to a plastic base, in part to boost the spacecraft's final altitude. The fuel reacts with liquid nitrous oxide to propel the rocket. Hart said that the fuel and oxidizer tanks were found intact. In 2007, an explosion during a ground-based test with nitrous oxide killed three workers at Scaled Composites, the Mojave-based company that designed and built *SpaceShipTwo*.

*SpaceShipTwo* is a larger, eight-seat version of *SpaceShipOne*, the two-seater that won the US\$10-million Ansari X Prize in 2004 on becoming the first private vehicle to repeatedly reach an altitude of 100 kilometres.

Virgin Galactic had been selling seats on future flights, with up to 5 minutes of weightlessness, for \$250,000 each. Its customers include not only celebrities but also scientists

who want to use it for microgravity studies. A second *SpaceShipTwo* is about 60% built. Meanwhile, a competing company, XCOR Aerospace of Mojave, is building a piloted space plane that would take tourists up one at a time.

The accident underscores the complexities of private spaceflight, where engineering systems are designed from scratch and tested in very public view. Many compare the endeavour to the early days of aviation, when aeroplane companies crashed time and again as they tried to commercialize air transportation. "To some degree, we are still in the infancy of spaceflight," says Scott Hubbard, director of Stanford University's centre of excellence for commercial space transportation in Palo Alto, California.

Commercial spaceflight will probably survive, analysts say, but only if the public is as willing to accept the risks as are aerospace experts. "I don't mind if it takes some time to develop," says Alan Stern, a planetary scientist at the Southwest Research Institute in Boulder, Colorado, who has signed up for a research seat on *SpaceShipTwo*. "I'm quite convinced that commercial suborbital flight will be safe." NASA is building a new launch system for

astronauts, but it is space companies that are coming up with rockets and spacecraft designs.

As the space-tourism industry tries to assess its future, another area of commercial spaceflight is likely to keep forging ahead. Last week's failed Antares mission was part of a series of cargo flights to the International Space Station organized by Orbital Sciences of Dulles, Virginia, in partnership with NASA. An inquiry led by Orbital aims to find out why the rocket lifted briefly off its launch pad at Wallops Island, Virginia, before exploding and falling to the ground.

The next planned step for private companies will be flying US astronauts, as well as cargo, to the space station. In September, SpaceX of Hawthorne, California, and Boeing of Houston, Texas, won a contract to begin flying astronauts by 2017. Hubbard, who chairs a safety committee to review SpaceX's upcoming crewed flights, notes that Orbital is not in the running for those.

"Accidents with new technology are inevitable," says Johnson-Freese. "How they are handled is the true test of innovation and innovators." ■

## CONSERVATION

# Future of Great Barrier Reef divides scientists

*Marine-park management comes under scrutiny as conservationists descend on Australia.*

BY DANIEL CRESSEY

The health of the world's most famous swathe of ocean real estate — the Great Barrier Reef Marine Park — will take centre stage next week as conservationists from around the world head to Sydney, Australia, for a once-in-a-decade meeting on ecosystem management. The park faces challenges, but scientists disagree over how endangered it is and how well it is being managed. Climate change further complicates the picture.

Every ten years, the International Union for Conservation of Nature hosts the World Parks Congress to decide how to use parks to promote conservation (see page 28). On 12–19 November, a particular focus will be how to enhance and expand marine parks (see 'Marine parks on trial'). Yet the Great Barrier Reef, which was once held up as a shining example of ecosystem management, has run into trouble.

Lying off the eastern coast of Australia, the park covers an area of ocean approximately the size of Germany, encompasses 3,000 coral-reef

systems and is the largest 'living structure' on Earth. It is managed by the Great Barrier Reef Marine Park Authority (GBRMPA), which has divided it into zones that impose different restrictions on activities, such as scuba diving or fishing.

This year, protests escalated over a proposed port expansion that would have dumped dredge material within the park boundaries. The plan was abandoned but the United Nations Educational, Scientific & Cultural Organization (UNESCO) will next year be deciding whether damage done to the park through degradation and development means that it should be included on the List of World Heritage in Danger. In August, the GBRMPA itself published a report warning that "the overall outlook for the Great Barrier Reef is poor and getting worse."

Among the evidence for problems is a much-cited 2012 study showing that coral cover had

halved between 1985 and 2012 (G. De'ath *et al.* *Proc. Natl Acad. Sci. USA* **109**, 17995–17999; 2012). The report placed much of the blame on cyclones and unusually large swarms of crown-of-thorns starfish (*Acanthaster planci*), which eat reef-building corals.

Some believe that much of the damage is temporary. Aaron MacNeil, who is studying the Great Barrier Reef at the Australian Institute of Marine Science in Townsville, points out that two huge cyclones — Hamish in 2009 and Yasi in 2011 — hit the reef in a particular way, producing a combined battering expected just once every 600 years. "I think in general the Great Barrier Reef is in pretty good shape but it has had a rough few years of storm activity that have left coral cover unusually low," he says.

Others, however, say that things are worse than they seem. Marine palaeoecologist John Pandolfi at the University of Queensland in Brisbane has been using sediment cores and other methods to reconstruct the reef's history over the past 1,200 years. "I am afraid that if you compare the current state of the reef to the

**"The overall outlook for the Great Barrier Reef is poor and getting worse."**





The Great Barrier Reef is struggling to cope with the impacts of climate change and development.

kinds of long timescales that my team looks at, then things are even worse than what you might have heard," he says.

Pandolfi's team has documented declines in *Acropora* corals, which are vital to the reef framework, that date back to the 1920s (G. Roff *et al. Proc. R. Soc. B.* **280**, 20122100; 2013). The losses are probably linked to changes in agriculture that were brought in by European settlers and affected water quality and damaged the reefs. Current reports therefore may underestimate declines in reef quality because they are often based on comparisons to a degraded state of the reef, rather than its truly pristine state, an issue known as shifting baselines. The GBRMPA is using Pandolfi's work to try and address this problem.

Modern pressures include the effects of nearby land development, such as run-off fertilizer from farming. Russell Reichelt, the GBRMPA's chairman and chief executive, says that the threat from the proposed dredge dumping was overblown, but that the GBRMPA will encourage government and local businesses to adopt a policy whereby their activities have a positive

net impact on the reef, not just a neutral one, as is the case now. It is also introducing targets for maintaining habitats and species as well as systems for assessing cumulative impacts.

Marine scientist Bob Kearney at the University of Canberra says that designating the reef as a marine park has fostered an "inappropriate" focus on fishing, given the bigger threat posed by climate change. The reef is highly sensitive to changing temperatures and ocean acidification, but global action is needed to tackle the carbon emissions that are the root cause of these issues. Similarly, last week, the Australian Academy of Science criticized a multimillion-dollar reef sustainability plan drawn up by the Australian and Queensland governments because it "fails to effectively address" any of the major pressures on the reef.

MacNeil is more optimistic, thanks to the current collaborative approach between the government, universities and the private sector to solving the reef's problems. "By working together I think we're in a better position to understand and address threats to the Great Barrier Reef than ever before," he says. ■

## MARINE PARKS ON TRIAL

### Indonesian corals in controlled test

The effect of marine park areas (MPAs) — parts of the ocean that are protected and managed for conservation — is difficult to tease out, not least because ocean ecosystems are affected by so many variables. Now, a study akin to a randomized controlled trial is aiming to do just that in the remote Bird's Head Seascape in Indonesia.

Sitting in the 'coral triangle' north of Australia, the seascape contains more than 2,000 islands, many of which have remained undeveloped. Gabby Ahmadi,

a marine scientist at the conservation organization WWF, has been searching for areas that can be used as controls for seven newly created MPAs within the Bird's Head.

Distance from fishing markets and exposure to waves, for example, must be matched between a park and its control. Ahmadi's team will then measure factors such as fish biomass to determine whether the park designation is making a difference. "We really need to get to the question of are these MPAs working, and, when they are working, why," she says. **D.C.**

## FUNDING

# Basket of gifts for German science

*Despite looming recession, research gets a windfall.*

BY QUIRIN SCHIERMEIER

Germany's economy may be shaky, but its science is in rude health. On 30 October, ministers from the federal government and Germany's 16 states pledged €25.3 billion (US\$31.6 billion) over the next 6 years to support special programmes for research and higher education.

With a research-and-development expenditure of about €100 billion per year — some two-thirds of which is corporate money — Germany is the fourth-highest science spender worldwide, behind the United States, China and Japan. Public funding for special programmes remained strong even after the global financial crisis of 2008. Now the windfall looks set to continue for the country's universities and large science organizations, despite drops in exports and industrial production that suggest that the economy could tip into recession.

The bulk of the money is for universities. Currently, some 2.6 million students are enrolled at Germany's roughly 300 universities — around 400,000 more than in 2005. First-year enrolment is projected to grow by 360,000 or so by 2020. To cope, the federal government plans to continue a pact with the states, launched in 2007, to jointly contribute nearly €20 billion by 2020.

Ministers have also agreed to continue a deal that guarantees 4 non-university research organizations, including the Max Planck Society, annual budget increases of 3% until 2020. Since 2005, the combined budget of the four organizations, which run a total of 254 institutes and large research centres, has risen from €5.2 billion to €7.9 billion. The rises are not earmarked for particular fields, but national priorities include energy, specifically renewables, and health, owing to the ageing population.

The federal government also plans to boost the budget of its main grant-giving agency for university research by 5% next year to €2.95 billion, and by 3% annually from 2016 on. And ministers agreed in principle that the German Excellence Initiative, a €4.6-billion university competition for top-up funds launched in 2006, may continue beyond its initial expiration year of 2017. ■



The reality of the Ebola outbreak is not reflected by model projections of high case numbers.

## EPIDEMIOLOGY

# Models overestimate Ebola cases

*Rate of infection in Liberia seems to plateau, raising questions over the usefulness of models in an outbreak.*

BY DECLAN BUTLER

The Ebola outbreak in West Africa has infected at least 13,567 people and killed 4,951, according to figures released on 31 October by the World Health Organization (WHO). Now, in a rare encouraging sign, the number of new cases in Liberia seems to be flattening after months of exponential growth. Scientists say it is too soon to declare that the disease is in retreat: case data are often unreliable, and Ebola can be quick to resurge. But it is clear that mathematical models have failed to accurately project the outbreak's course.

Researchers are now struggling to understand whether reports of empty beds at treatment centres and declining burial numbers are signs that fewer people are developing Ebola, or whether cases and deaths are going unrecorded. In Liberia's capital, Monrovia, just

80 of 250 beds were filled at the Médecins Sans Frontières (MSF) centre last week. But Fasil Tezera, who heads MSF's Liberia mission, is cautious: "The present epidemic is unpredictable," he says.

Epidemiologists normally use mathematical models to estimate the trajectory of an outbreak, and to estimate where and how to direct scarce medical resources. But for the current crisis, on-the-ground data contradict the projections of published models, says Neil Ferguson, an epidemiologist at Imperial College London, and a member of the WHO's multidisciplinary Ebola Response Team.

On 7 October, for example, modeller Alessandro Vespignani of Northeastern University in Boston, Massachusetts, and his collaborators predicted that Liberia would see 6,900–34,400 cases by 24 October, and 9,400–47,000 by 31 October. But the WHO

put the number of reported cases in the country at just 6,535 as of 25 October.

Vespignani says that his model was a worst-case scenario, in which exponential growth of cases continued and containment measures were ineffective. But he and other modellers are also handicapped by incomplete and unreliable data on Ebola epidemiology, especially in the hardest-hit areas. And they have little empirical data on how disease-control measures quantitatively affect Ebola transmission, says ecologist Nick Golding, who studies the spatial distribution of disease at the University of Oxford, UK. Models "are fitted to pretty poor-quality data on case counts, and essentially no data on interventions", he says, making it difficult to generate accurate projections.

Two more-complex models published last month attempted to tease out the effects of various control measures. But their outcomes also do not square with the most recent Liberia data (J. A. Lewnard *et al. Lancet Infect. Dis.* <http://doi.org/wn9>; 2014, and A. Pandey *et al. Science* <http://doi.org/wts>; 2014). That does not surprise Alison Galvani, an epidemiologist at Yale University in New Haven, Connecticut, and an author of both studies. "Epidemics are moving targets," she says, adding that her model projections are at best a preliminary outline for public-health intervention. Because the model projections can be easily misunderstood, Ferguson says that modellers "really need to think carefully about what we really know about Ebola transmission and the impact of different interventions, and do our best to communicate the many uncertainties".

In the meantime, Bruce Aylward, a WHO assistant director-general who is coordinating the agency's Ebola efforts, is "terrified" that any plateau in new cases will be misinterpreted as meaning that the problem is going away. There is still a need to greatly increase the resources available to treat infected people and prevent new cases, Aylward says.

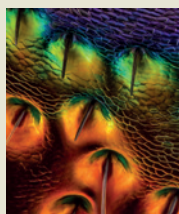
But if the slowing rate of infection in Liberia is confirmed, it could suggest that even moderate levels of public-health intervention can pay off, says Golding. For the current Ebola outbreak, the average number of new cases spawned by an infected individual — 1.2–2.2 — is much lower than that of many other communicable diseases, such as measles (which can spread to between 12 and 18 people per case). As Ebola prevention measures push down this figure, the disease becomes easier to control; when it dips below 1, virus spread stops completely.

Until the West African outbreak is extinguished, there is a real risk that the disease will resurge in areas where it has been stamped out — or even cover new ground. A stark reminder of this came in the past two weeks: a two-year-old girl with Ebola travelled hundreds of kilometres from Guinea to Mali on a bus — raising concerns that the many people she came into contact with could spark outbreaks in Mali. ■

DANIEL BERHULAK/NYT/REDUX/EVINE

  
**MORE  
ONLINE**

## IMAGES OF THE MONTH



The most memorable science images from October [go.nature.com/6cmvhh](http://go.nature.com/6cmvhh)

## TOP NEWS

- Pet trade could spread amphibian fungus [go.nature.com/ivngd3](http://go.nature.com/ivngd3)
- A 'universal' index ranks heatwaves [go.nature.com/dcz92e](http://go.nature.com/dcz92e)
- Unschooled Mayans have good grasp of probabilities [go.nature.com/oekidl](http://go.nature.com/oekidl)

CHARLES KREBS/COURTESY OF NIKON SMALL WORLD



## POLITICS

# Lobbying sways NIH grants

*Pressure on lawmakers from patient-advocacy groups has shaped agency spending on rare-disease research.*

BY SARA REARDON

Advocates for patients with rare diseases spend millions of dollars lobbying the US Congress each year — and it is money well spent, an economic analysis has found. Between 1998 and 2008, such efforts helped to increase new funding for rare-disease programmes by 3–15% each year at the US National Institutes of Health (NIH), according to a report to be published in *Management Science* (D. Hegde and B. N. Sampat *Mgmt Sci.* <http://doi.org/fzs2vx>; 2014).

The effect of this growth on the NIH's total budget (US\$30 billion in fiscal year 2014) is small; targeted grants accounted for just 10–20% of the agency's annual grant-making during the study period. But the analysis highlights the fine line that the NIH must tread when choosing diseases to prioritize: maintaining the peer-review process by which it awards grants, but not ignoring the wishes of the lawmakers who control its budget.

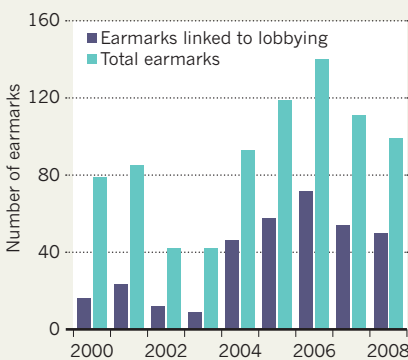
Congress curbed the power of lobbyists through a 2010 ban on setting aside money in bills for specific projects — known as earmarks. But lobby groups have shifted strategies. They seek to steer funds using 'soft' earmarks: language in spending bills that encourages or urges an agency to perform some action, such as funding Alzheimer's research, rather than setting aside funds for it (see 'Research rewards').

When economists Deepak Hegde of New York University and Bhaven Sampat of Columbia University in New York City examined the text of congressional reports on the NIH budget for mentions of 955 rare diseases, they found an average of 84 soft earmarks a year for these conditions. The true impact of lobbying on the NIH budget is likely to be even larger than their estimates, they say, because the study examines just a small slice of the agency's research portfolio that does not include diseases such as cancer.

Rachel Best, a sociologist at the University of Michigan in Ann Arbor, says that lobbying provides a way for taxpayers to communicate their priorities to the NIH — and it pays. She analysed funding for 53 diseases over the course of 19 years and found that each \$1,000 spent on lobbying for a condition correlated with a \$25,000 increase in funding the following year

## RESEARCH REWARDS

Lobbying by patient-advocacy groups is linked to a rise in 'earmarks' in bills funding rare-disease research at the US National Institutes of Health.



(*R. Best Am. Sociol. Rev.* 77, 780–803; 2012).

This could be problematic if it starts to skew funding towards research that has little scientific merit or comparatively small public-health benefit, Best adds. She notes that the diseases that affect the most people do not necessarily get the strongest advocacy: she has found that illnesses that primarily affect women and minorities, and stigmatized conditions such as sexually transmitted diseases, tend to receive less lobbying support than other conditions.

Jeremy Berg, a biochemist at the University of Pittsburgh in Pennsylvania and a former director of the National Institute of General Medical Sciences, says that he is neither surprised nor concerned by the findings. If Congress uses a soft earmark to encourage research on a particular topic, he says, "it's something NIH takes seriously" — but does not feel compelled to act on. "In my experience, it's not regarded as a command, but input into what to do."

And Pierre Azoulay, an economist at the Massachusetts Institute of Technology in Cambridge, says that it would be interesting to determine whether targeted grant programmes for specific diseases chosen by the NIH yield better results than grants proposed by individual researchers. "To know whether to be exercised about this, we should look at how scientifically fruitful those different grants are," he says. Hegde says that his group is currently doing just this for all NIH research funding. ■

SOURCE: D. HEGDE & B. N. SAMPAT *Mgmt Sci.* <http://doi.org/fzs2vx> (2014)



Mark Armitage was fired from his job as a lab technician at California State University, Northridge.

SCIENCE AND RELIGION

# University sued by creationist

*Microscopist's wrongful-dismissal case faces long odds.*

BY CHRISTOPHER KEMP

In May 2012, Mark Armitage made a discovery that he had dreamed of for years. While digging in Montana, he uncovered one of the largest triceratops horns ever found in the Hell Creek Formation, a legendary stack of fossil-bearing rocks that date to the last days of the dinosaurs. Armitage drove the horn back home to Los Angeles, California, where his microscopic examination revealed that it contained not only fossilized bone but also preserved layers of soft tissue. "They were brown, stretchy sheets. I was shocked to see anything that was that pliable," he says.

In February 2013, he published his findings in *Acta Histochemica*, a journal of cell and tissue research (M. H. Armitage and K. L. Anderson *Acta Histochem.* **115**, 603–608; 2013). Two weeks later, he was fired from his job at California State University, Northridge (CSUN), where he managed the biology department's electron and confocal microscopy suite.

Now he is embroiled in a long-shot legal fight to get his job back. In July, his lawyers filed a wrongful-termination suit claiming that religious intolerance motivated the dismissal: as a young-Earth creationist, Armitage says that finding soft tissue in the fossil supports his belief that such specimens date to the time of the biblical flood, which he puts at about 4,000 years ago.

The suit alleges that faculty members hostile to Armitage had him fired because they could not stand working with a creationist who had been published in a legitimate scientific

journal. He and his attorneys at the Pacific Justice Institute, a conservative legal organization based in Sacramento, California, that focuses on religious and family issues, have repeatedly made that claim in the press. But specialists in US labour law suggest that his claim of religious intolerance might have difficulty standing up if the case goes to trial.

In recent years, a schoolteacher, academic and NASA employee who were creationists have claimed that they were fired unjustly for their religious beliefs. (None were reinstated.) But what makes this case different is that Armitage managed to survive for years in a mainstream academic institution and to publish research in a respected peer-reviewed journal.

Armitage acknowledges that he did that by keeping his views on the age of the fossil out of the paper. Written with biologist Kevin Lee Anderson of Arkansas State University-Beebe, the study simply reported that the horn was found in Hell Creek (which has a well-accepted age of 65 million to 70 million years). "It was just morphology," says Mary Schweitzer, a palaeontologist at North Carolina State University at Raleigh who reviewed the work before publication, and made the first discovery of soft tissue in dinosaur bones in 2005. "It was fine."

Creationists often appeal to soft-tissue preservation as evidence that dinosaur fossils are thousands rather than millions of years old, says palaeontologist Jack Horner of the Museum of the Rockies in Bozeman, Montana. "Science is about building hypotheses and then attempting to falsify them," he says. "Creation science or any

kind of pseudoscience is just the opposite. It is coming up with an idea or a notion or anything else and finding evidence to support it."

According to his lawsuit, Armitage never tried to conceal his beliefs from his employer. The filing says that when he was interviewed for his job at CSUN in November 2009, he did not hide that he holds degrees from the Christian-fundamentalist Liberty University in Lynchburg, Virginia, and the Institute for Creation Research, previously in San Diego, California. In an interview with *Nature*, Armitage said that he was also equally open about his roughly 30 technical articles on microscopy and his 2008 self-published book *Jesus is like my Scanning Electron Microscope*.

The lawsuit also claims that Armitage excelled in his job, receiving numerous letters of commendation. "I'm not a microscopist but as far as I could tell, Armitage was a good one," Paul Wilson, a biologist at CSUN, told *Nature*.

CSUN declined to comment on Armitage's performance or its reasons for ending his employment. However, Jeffrey Noblitt, associate vice-president of marketing and communications at CSUN, did stress in an email that Armitage's position had been "temporary".

Armitage freely admits that he often engaged students in conversations, giving his opinion on issues such as the age of the remarkably well-preserved cells in the triceratops horn. "To me, the obvious conclusion is they're young. They can't be 68 million years old," he says.

In terms of getting his job back, those conversations might be Armitage's undoing. US anti-discrimination laws require employers to reasonably accommodate an employee's beliefs or religious practices, unless doing so would cause 'undue hardship' to the employer, says Justine Lisser, a spokesperson for the US Equal Employment Opportunity Commission.

If Armitage made his living bending metal in a machine shop, an employer would find it difficult to show how his views caused undue hardship, she says. But in an academic setting, telling biology or palaeontology students that life began only a few thousand years ago more clearly undermines the institution's goals. "It would be an easier showing of undue hardship," says Lisser, "because it's more related to the essence of what the person is doing." ■

## CORRECTION


In the News story 'Geneticists tap human knockouts' (*Nature* **514**, 548; 2014), the team investigating knockouts found some 200,000 variations that knocked out genes (not 150,000 genes, as stated). In addition, the team that studied 36,000 Finnish people was not led by Daniel MacArthur, he was just a member of the group. Although Marfan syndrome can cause serious heart problems, it rarely results in sudden heart failure. Finally, Bing Yu is a woman not a man.



# CENTRAL EUROPE UP CLOSE

*In the 25 years since the collapse of communism, the countries of central and Eastern Europe have each carved their own identity in science.*

BY ALISON ABBOTT & QUIRIN SCHIERMEIER

	mag 𐀀	WD	HFW	det	mode	HV	curr	25 µm	
	1 200 x	3.9 mm	106 µm	TLD	SE	10.00 kV	0.40 nA	Helios NanoLab 450HP	

In a lab so new that it still smells of fresh paint, Katarzyna Komorowska expertly handles what looks like a futuristic coffee machine. It is actually an advanced scanning electron microscope with the power to manipulate delicate samples and visualize minute details — one of several impressive-looking machines in Komorowska's lab in the city of Wrocław in southwest Poland. Komorowska turns on the device's ion beam. Minutes later, a screen shows the razor-sharp image of a bearded dwarf clutching a graphene molecule that she has just engraved on a grain of sand.

The etched sand is a historical reminder as well as a technological feat. The dwarf became an unlikely symbol of the 1980s protest movement that grew in Wrocław against Poland's ruling communist regime. It is now something of a city mascot: Wrocław hosts more than 300 dwarf statues, and visitors can track them down using a brochure and app. The fact that the dwarf can be engraved on a grain of sand in seconds also symbolizes the formidable

efforts that this city is making to become a science hub in central Europe. Since 2007, more than €200 million (US\$250 million) in European Union (EU) funds have helped to turn Wrocław's abandoned military hospital into a campus dedicated to academic and commercial science — just one part of Poland's high-flying ambitions for science as a whole.

Change has swept through central and Eastern Europe since the collapse of communism there 25 years ago. The revolution was quick and unforeseen. For a few months in 1989, protests swelled behind the Iron Curtain, the political barrier that since the end of the Second World War had isolated communist central and Eastern European countries from the West. Then, on 9 November that year, the East German government opened the Berlin Wall and first a trickle — then a flood — of East and West Germans began to scale the barrier, delirious with joy. A year later, Germany had been reunified and almost every other former communist country in the region had instituted

a democratic government.

Researchers shared in the elation: the fall of the Iron Curtain brought them personal and intellectual freedom. But it came with a host of new problems. During the 45-year communist rule, research institutions from the Baltic to the Balkans had been academically isolated and unable to compete with the rest of the world. Now they were suddenly being judged by international standards, and their science looked hopelessly out of date. For many, political change also brought poverty, as economies collapsed. Pitiably low salaries, lack of funding and antiquated labs prompted swathes of scientists to go west or seek careers outside academia. Those who stayed relied almost exclusively on foreign aid. "After the Iron Curtain had come down, science and higher-education institutes were thrown into turmoil," says Liviu Mattei, pro-rector of the Central European University in Budapest. "Few

**Scanning electron microscope image of a grain of sand engraved at EIT+ in Wrocław, Poland.**

KATARZYNA KOMOROWSKA/WROCLAW RES. CENTRE EIT+



places in the world have gone through such rapid and brutal changes.”

Twenty-five years on, researchers find themselves in a more stable scientific landscape. The economic decline of the 1990s has mostly ended, and in the past decade some countries have enjoyed a marked economic upswing that has allowed governments to inject money into science. Membership of the EU has been a major driver of change. In 2004, the union welcomed eight former communist countries, including Poland, Estonia and Hungary. Romania and Bulgaria followed in 2007, and Croatia in 2014. One EU citizen in five now lives in one of these new member states.

These relatively poor countries have enjoyed huge financial injections from EU structural funds, which are designed to narrow economic and social disparities between European regions and are distributed by each country's government. In the 2007–13 financial period, Brussels invested a staggering €170 billion in cohesion and regional development in the new member states, and more than €20 billion of this was earmarked for science and innovation. Most countries have also created funding agencies that allocate grants on a strictly competitive basis. “Scientists had to learn that performance is now the sole basis of getting funded and published,” says Franci Demšar, director of the Slovenian Research Agency in Ljubljana. “It has been a difficult process, but it has greatly improved science produced in this part of the world.”

But within central and Eastern Europe, different nations have followed starkly different trajectories in science, as a spotlight on three countries in the region reveals (see ‘Science in the new Europe’). Poland hosted relatively little research until recent years, but the nation is now becoming a political and economic powerhouse in the region and is rapidly expanding in science. Estonia, a small country on Europe's northern fringes, reformed its research system early on and is now reaping the benefits. Hungary, by contrast, maintained some scientific strengths during the communist era, but a lack of investment is now putting that legacy at risk.

This means that when it comes to science, central and Eastern European countries — so similar to each other in their communist days — are growing steadily apart. What is more, almost all are still fighting a brain drain to the West. “The talent for science is all there,” says Lars Walløe, a physiologist at the University of Oslo and former president of the Academy of Europe. “Now the conditions and institutions in the region need to develop in such a way that the best minds will find it worthwhile to stay.”

#### POLAND: POCKETS OF EXCELLENCE

Poland has embraced science like few other countries in the region, as is evident on the Wrocław campus with its dwarf-engraving machine. The campus has the air of a sprawling, half-built start-up company. Extensive lab spaces are still under construction and will be

opened to scientists and entrepreneurs next year. On a rainy September morning, labs and meeting rooms are buzzing with scientists testing equipment and discussing results.

The campus is called EIT+, to echo the Budapest-based European Institute of Innovation and Technology, an EU effort to create a network of research powerhouses. Scientists at EIT+ pursue independent research in subjects including nanotechnology, materials science and biotechnology. But the campus operates as a limited company that aims to provide industry with research and services such as microscopy and crystallography, at a profit. “Twenty years ago the kind of things we're doing would have been unthinkable,” says Jerzy Langer, head of EIT+ and a former Polish deputy science minister. “The time has long gone that scientists here could say ‘Sorry, I can't do this or that, I haven't got the money and the tools.’”

Komorowska joined EIT+ in 2012. She had trained in Wrocław, but left the country for postdoctoral work in France and Belgium, with no plans to return. She changed her mind

## AFTER THE IRON CURTAIN HAD COME DOWN, SCIENCE WAS THROWN INTO TURMOIL.

when she got a job offer from the newly created Wrocław Research Centre, part of EIT+. She now leads the centre's laboratory of nanotechnology and semiconductor structures, where she is developing an automated system for analysing the mineral and metal content of rocks — data of use to the massive Polish mining industry. Some 35 million złoty (US\$10.5 million) are being spent on furnishing the lab with the latest electron microscopes for characterizing and observing materials. “The conditions to do science in Poland have improved enormously now that we have the same equipment as most people in the West,” says Komorowska.

That was not true in the communist era, when the country operated just a few basic-research institutes. The situation began to change in 1990, when Lech Wałęsa, leader of the trade union Solidarity, took over as Polish president and began to modernize the country. Poland went through a painful transition when democracy and a market-based economy arrived — and science was shaken to the core. Seeking more lucrative opportunities, thousands of researchers went into business or left to pursue academic careers abroad. What remained of the communist research base was jealously guarded by an increasingly inward-looking group of ageing academics and produced little in terms of internationally competitive science.

The situation really turned around when Poland joined the EU in 2004. With 38.5 million inhabitants, the country is by far the most populous member state in the region, and also receives the most EU structural funds. That money has helped to fuel its remarkable economic growth, which has been outpacing that of most other European countries since 2008. Science has ridden on the coat-tails of the country's thriving economy, and the government has recognized that research is an important route to further growth. Domestic funding has doubled over the past five years, although overall science expenditure is still very low compared to other places in Europe, at less than 1% of gross domestic product (GDP).

Outside Wrocław, the picture is not entirely rosy. Polish science still has a workforce shortage, with fewer than 4 researchers per 1,000 people in the labour force — well below the EU average of just under 7. Scientists acknowledged that many of the country's research institutes — particularly the 80 run by the Polish Academy of Sciences — are reluctant to reform. What is more, says Langer, at some university departments a spirit of obedience lingers from the communist days and tends to stifle creativity. “Many students are too shy,” he says. “They think that ideas are no good if the authorities haven't rubber-stamped them.”

To address the problem, Poland's National Centre for Research and Development now runs a programme designed to bring back early-career Polish scientists who trained abroad, and to attract foreign scientists, by offering them up to 1.2 million złoty to start an independent group. (Securing this funding was another reason that Komorowska was persuaded to return.) And there is a wave of fresh talent on the way. The number of science graduates from Polish universities has more than doubled over the past decade, and overall student numbers have more than quadrupled since 1990. Every tenth student in the EU is now Polish. “This is most encouraging,” says Langer.

Scientists in Poland still want to see better standards of science education, and more innovative institutes such as EIT+. “There is sufficient money available now to support promising ideas, but we need to create institutional environments in which science can flower,” says Janusz Bujnicki, a molecular biologist at the International Institute of Molecular and Cell Biology in Warsaw. “That task has only just begun.”

#### ESTONIA: SMALL AND FOCUSED

Andres Metspalu, who runs one of the world's most sought-after banks of human DNA, says that his life began at 40. That was in 1991, the year Estonia declared its independence from the Soviet Union and began its bumpy path towards a Western-style research base.

Before this, Metspalu's ambition to do world-class science had been constantly undermined by the absurdities of the isolationist Soviet system. That system had acknowledged his talent:



# SCIENCE IN THE NEW EUROPE

Since 2004, 11 former communist states have joined the European Union, gaining funding that has aided an economic upswing and increased investment in science. But they have followed different trajectories. Some are fully embracing modern science and are almost equal partners in Europe; others lag behind.

## PUBLICATION COUNT

Central and Eastern Europe produce about 4% of the world's publications, with Poland taking the lion's share. This is still dwarfed by Western powerhouses such as Germany and the United Kingdom.

UNITED KINGDOM  
130,313

GERMANY  
153,431

POLAND  
29,099

POLAND  
29,099

CZECH REPUBLIC  
13,563

SLOVENIA  
4,575

HUNGARY  
7,382

BULGARIA  
3,203

SLOVAKIA  
4,426

LITHUANIA  
2,232

ESTONIA  
1,949

LATVIA  
951

ROMANIA  
10,166

Poland has undergone sharp economic growth and is rapidly enlarging its research base.

Number of articles published in Elsevier's citation database Scopus in 2013

it selected him in 1981 as one of only 25 young scientists across the entire Soviet Union to spend a year training in a US laboratory after his doctorate. (He studied biochemistry at Columbia University in New York, then at Yale University in New Haven, Connecticut.) But his young family had to stay in Tartu, Estonia, so that the authorities could be confident that he would not defect. And after that, he was forbidden further travel until Mikhail Gorbachev swept to power in the Soviet Union in 1985. Over the next few years, Metspalu watched the Soviet Union fall apart.

With a population of just 1.3 million people, Estonia found change easier than some larger countries. This, and some good political decision-making, helped to make the country one of the first in the region to reverse its economic decline. The government was prudent enough to stimulate business-orientated research and entrepreneurship early on. Newly created research centres and technology parks in Tartu and the capital, Tallinn, became the nuclei for a remarkable scientific upswing.

After the human genome was sequenced at the turn of the millennium, Metspalu, who worked at the University of Tartu, saw a scientific opportunity. Thanks to support from Estonia's EU structural funds, he was able to launch a project to recruit individuals to donate their genetic and health data to a national biobank. In other countries, such as Iceland, similar efforts to collect personal

information en masse were met with suspicion. But that was not the case among the newly liberated and optimistic people of Estonia, many of whom were happy to sign up. The biobank now includes genetic and health information on 5% of the country's adult population and is a valuable international resource in studies that require very large numbers of people to identify risk genes associated with common diseases, including obesity<sup>1</sup>

**TWENTY-FIVE YEARS AGO  
I'D HARDLY HAVE IMAGINED  
CONTRIBUTING TO A  
BIOMEDICAL REVOLUTION.**

and schizophrenia<sup>2</sup>. The database has been "extremely helpful", says Michael O'Donovan, a psychiatrist at Cardiff University, UK, who was involved in the schizophrenia study.

As well as biotechnology, the Estonian government is focusing scientific investment on disciplines such as materials science and informatics. The chemistry department at the University of Tartu, for example, is prominent in the area of superacids and superbases, useful in the development of batteries for electric

cars, and has collaborations with car manufacturers in several countries. Estonia has steadily increased its investment in research and development, from 0.72% of the GDP in 2002 to 2.18% in 2012 — the second highest in central and Eastern Europe, after Slovenia.

The country is keen to reap the rewards of its investments by using the biobank for more than just research. This year, the government formally pledged to support a project that would — in the next few years — link the repository with Estonia's centralized health database to allow physicians to support their diagnoses and therapies with individual genetic information. If all goes to plan, this will put Estonia among the world's front-runners in personalized medicine. "Twenty-five years ago I'd hardly have imagined being able to contribute to a biomedical revolution," says Metspalu.

## HUNGARY: SCIENCE ON A SHOESTRING

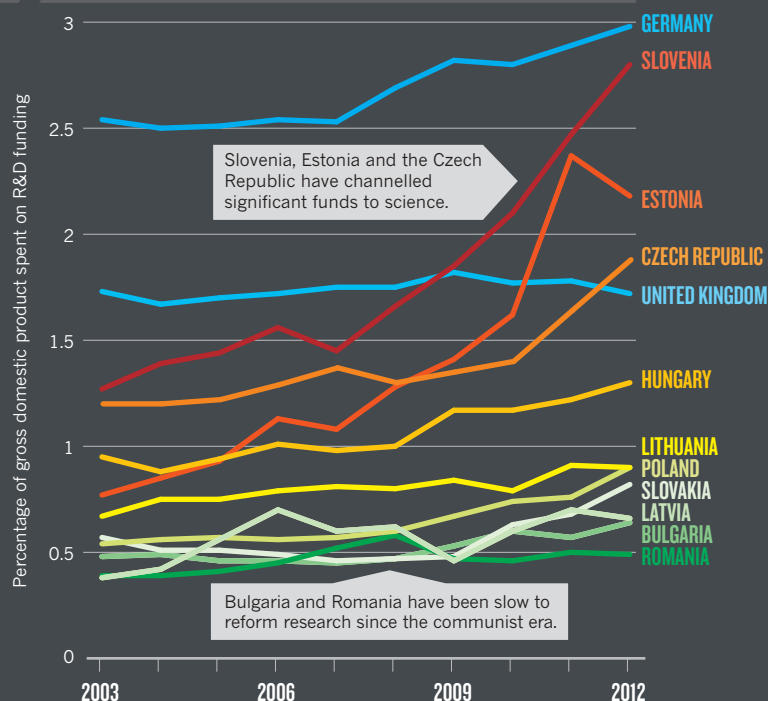
Last year, plant biologist Eva Kondorosi decided to pack up her life in Paris and take her European funding to the Hungarian Academy of Sciences (HAS) Biological Research Centre, where she had begun her research career in the late 1970s. But she found something odd: unlike many research institutes across central and Eastern Europe, this one seemed to be less exciting in the modern era than it had been in the communist one. "It has lost the vibrancy we enjoyed there in the 1970s and 80s," she says.

If that sounds paradoxical, it is because

SOURCE: PUBLICATIONS: SCOPUS; SPENDING: EUROSTAT; GRANTS: ERC

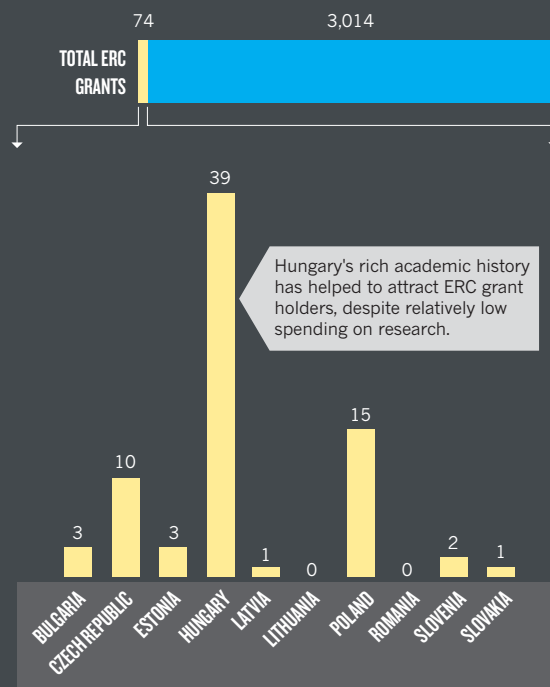
## RESEARCH SPENDING

Most central and Eastern European countries have increased total public and private investment in science over the past decade, and some have matched Western European spending.



## EUROPEAN RESEARCH COUNCIL GRANTS

Despite the expansion of science, central and Eastern European countries host only a small proportion of scientists with prestigious ERC grants.



science in Hungary is a bit of a paradox. During the communist years, research survived against the odds. It was separated from universities and centralized in institutes — run by the highly politicized HAS — that tended to appoint cronies and members of the ruling communist party to key research posts. At the end of the 1960s, however, the HAS made a bold decision to start afresh in biology by creating the Biological Research Centre in the southern city of Szeged, close to Hungary's southern border and away from the stifling politics of the capital. Staff members at the multidisciplinary centre were appointed on merit, not political status. The centre became a safe haven for intellectuals, and Kondorosi did a doctorate in plant sciences there. The country's borders were slightly more open than those of most of its communist counterparts, and throughout the 1980s researchers could gain permission to visit the West. This kept the science in Szeged cutting-edge, and the atmosphere buzzing.

After the fall of communism, lack of investment dampened enthusiasm, and many scientists left. Funding has never fully picked up. Hungary is one of only three countries to have reduced its public spending on research since 2007 (the others are Croatia and Bulgaria), and it dedicated just 8.5% of its 2007–13 EU structural funds to research — compared to Estonia's 20% and Poland's 14%. What is more, much of the structural funding has gone to companies, not academia.

Despite this, the country's rich academic legacy still attracts the best Hungarian scientists. Hungary has hosted more researchers with prestigious European Research Council grants — including Kondorosi — than any other former communist EU country. Kondorosi returned because the interdisciplinary organization of the Biological Research Centre offered her the opportunity to take the lessons she had learnt from studying plant–bacteria symbiosis and apply them to medicine. She and others have discovered antimicrobial activity in some of the bacteria that plants use for nitrogen-fixing<sup>3</sup>, for example.

Hungary's scientific culture has inspired international confidence. In 2012, CERN, the European laboratory for particle physics, built an advanced data centre close to Budapest. And Szeged is going to be home to one of the three nodes of the Extreme Light Infrastructure, a collaborative EU project to advance laser science (see *Nature* **489**, 351; 2012).

Scientists are hopeful that the climate for science is warming up. In June this year, the charismatic József Pálincás was appointed to the new position of government commissioner for science and innovation. Pálincás was formerly president of the HAS, where in 2011 he forced through reforms that streamlined the academy's 40 research units into 15 larger centres, and increased scientific competition for funding. In his new role he will be responsible for advising the government on science policy,

as well as coordinating the spending of the current round of research-related structural funds. In this round, he says, around 12% will be directed to research.

Scientists in Hungary and elsewhere have their eyes on the EU's Horizon 2020 programme, the €80 billion of research funding that started this year and will last until 2020. In a bid to widen participation, Brussels has launched a 'teaming' scheme that allows less potent member states to create or upgrade competitive research centres in partnership with leading institutions from other countries.

Financial and organizational aid will remain crucial in narrowing the gaps between countries, says Walløe. "Some places will always have more capacity of science than others," he says. "But every country should have at least one thing or the other that is really good. That is how science is organized in the United States — and that's what it should be like in Europe." ■ [SEE EDITORIAL P.7](#)

**Alison Abbott** is Nature's senior European correspondent and **Quirin Schiermeier** is Nature's German correspondent. Both are based in Munich.

1. Walters, R. G. et al. *Nature* **463**, 671–675 (2010).
2. Schizophrenia Working Group of the Psychiatric Genomics Consortium *Nature* **511**, 421–427 (2014).
3. Tiricz, H. et al. *Appl. Environ. Microbiol.* **79**, 6737–6746 (2013).

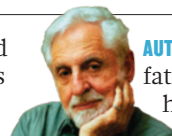


# COMMENT

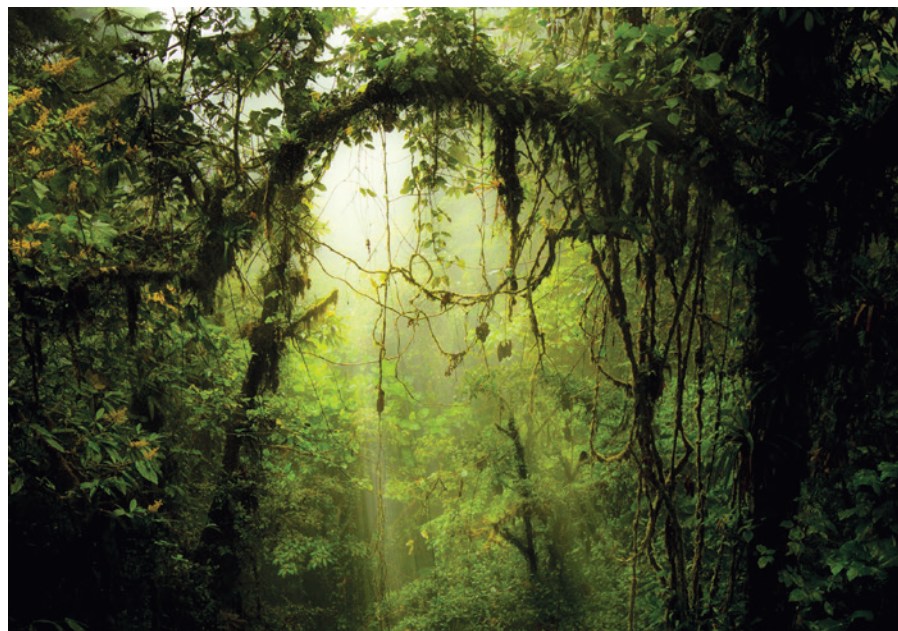
**CONSERVATION** Experts weigh in on priorities for the world's protected areas **p.28**

**ECONOMICS** Loss of natural and ecological capital leaves hole in national budgets **p.32**

**CHEMISTRY** The fame and shame of the elements that never were **p.34**



**AUTOBIOGRAPHY** Carl Djerassi, father of the Pill, reflects on his life and literature **p.36**



Conservation efforts risk getting snared in a tangle of aims.

## A call for inclusive conservation

**Heather Tallis, Jane Lubchenco and 238 co-signatories** petition for an end to the infighting that is stalling progress in protecting the planet.

An age-old conflict around a seemingly simple question has resurfaced: why do we conserve nature? Contention around this issue has come and gone many times, but in the past several years we believe that it has reappeared as an increasingly acrimonious debate between, in essence, those who argue that nature should be protected for its own sake (intrinsic value)<sup>1,2</sup> and those who argue that we must also save nature to help ourselves (instrumental value)<sup>3-5</sup>.

Champions of instrumental value contend, among other things, that protecting nature for its own sake alone has failed to stem the tide of species extinction, that conservation

should be open to partnering with business to effect the greatest change and that conservation support will be broadened by more directly considering other social objectives (such as food security or clean water). By contrast, advocates of intrinsic value assert that ethical arguments for conservation should be sufficient, that partnering with business is selling out to those who create the problem and that social considerations are already central to conservation.

Unfortunately, what began as a healthy debate has, in our opinion, descended into vitriolic, personal battles in universities, academic conferences, research stations,

conservation organizations and even the media<sup>6</sup>. We believe that this situation is stifling productive discourse, inhibiting funding and halting progress.

Adding to the problem, in our view, is the issue that this dispute has become dominated by only a few voices, nearly all of them men's. We see this as illustrative of the bigger issues of gender and cultural bias that also continue to hinder conservation.

The stakes? The future of conservation science, practice and policy. Conservation regularly encounters varied points of view and a range of values in the real world. To address and engage these views and values, we call for more-inclusive representation of scientists and practitioners in the charting of our field's future, and for a more-inclusive approach to conservation.

### EMBRACE DIVERSE VALUES AND VOICES

Women historically have been under-represented in environmental-science faculty positions and in conservation practice, as in most scientific fields. This disparity is changing globally, but at different rates: more slowly in Asia and more quickly in Latin America and the Caribbean, for example<sup>7</sup>. In the United States, more than half the leadership positions in conservation organizations are now held by women. And on the global stage, women currently hold top positions in many leading efforts, including the Intergovernmental Platform on Biodiversity and Ecosystem Services, the Future Earth science committee, and the International Union for Conservation of Nature. This progress makes the dearth of female voices in the debate about the premise of our profession all the more stark.

The signatories in agreement here — women and men from around the globe — support an equal role for women and for practitioners of diverse ethnicities and cultures in envisaging the future of conservation science and practice.

Together, we propose a unified and diverse conservation ethic; one that recognizes and accepts all values of nature, from intrinsic to instrumental, and welcomes all philosophies justifying nature protection and restoration, from ethical to economic, and from aesthetic to utilitarian. What we propose is not new. This diverse set of ethics has a long-standing history in modern conservation<sup>8</sup>. For ►

MIKE LANZETTA/GETTY

► example, more than 100 years ago, both intrinsic and instrumental values were used in the creation of Yellowstone National Park in Wyoming, and when Californians spurred the broader environmental movement in the United States by using economic studies of the value of birds alongside compelling speeches about the purity and grandeur of nature<sup>9</sup>.

These values need not be in opposition, although they do reflect the hard choices that conservation often faces. They can instead be matched to contexts in which each one best aligns with the values of the many audiences that we need to engage. Those on the side of intrinsic value will argue that by recognizing the many ways in which people benefit from nature, we cheapen nature and miss opportunities to save components of it that have little or no obvious value to people. This is a valid concern, and one of many reasons why we must continue to uphold intrinsic values to audiences who share those values, or may be inspired towards them. However, instrumental values will remain more powerful for other audiences, and should be used in the many contexts where broadening support for conservation is essential<sup>4</sup>.

Clearly, all values will not be equally served in every context. Approaching conservation problems with representative perspectives and a broad base of respect, trust, pragmatism and shared understanding will more quickly and effectively advance our shared vision of a thriving planet. Prominent institutions already embrace multiple voices and values. For example, the field's signature international

treaty, the Convention on Biological Diversity, calls for the conservation of biodiversity, and for the sustainable use and equitable sharing of its benefits. Some countries leading in this area, such as Mexico, Costa Rica and Colombia, have followed suit, capturing these joint interests in their own governing language.

#### PRACTICAL ACTION

What now? Academic training of conservation scientists should more accurately portray the rich, global history of the field, introducing students to the diverse ways in which nature has been valued and conserved for centuries. More forums at conferences, in journals and on social media are needed to elevate the voices of scientists and practitioners from under-represented genders, cultures and contexts. Conservation organizations and scientists can embrace all plausible conservation actors, from corporations to governmental agencies, faith-based organizations and interested individuals, and advance conservation efforts when they can benefit people and when there is no obvious human-centric goal.

These efforts must be underpinned by a stronger focus on synthesizing and expanding the evidence base that can identify what works and what fails in conservation so that we can move from philosophical debates to rigorous assessments of the effectiveness of actions. And we must encourage the full breadth of conservation scientists and practitioners to engage with the media so that coverage reflects the true range of opinion

(for example, the 240 co-signatories listed are ready for interview) rather than the polarized voices of a few. To add your name to this petition, visit [diverseconservation.org](http://diverseconservation.org).

It is time to re-focus the field of conservation on advancing and sharing knowledge in all relevant disciplines and contexts, and testing hypotheses based on observations, experiments and models<sup>10</sup>. We call for an end to the fighting. We call for a conservation ethic that is diverse in its acceptance of genders, cultures, ages and values. ■

**Heather Tallis** is lead scientist at the Nature Conservancy in Santa Cruz, California, USA. **Jane Lubchenco** is professor of marine biology and of zoology at Oregon State University in Corvallis, Oregon, USA. e-mail: [htallis@tnc.org](mailto:htallis@tnc.org)

1. Gudynas, E. in *La Naturaleza con Derechos: De la Filosofía a la Política* 239–258 (AbyaYala, Universidad Politécnica Salesiana, 2011).
2. Soulé, M. *Cons. Biol.* **27**, 895–897 (2013).
3. Reid, W. V. et al. *Nature* **443**, 749 (2006).
4. Kareiva, P. & Marvier, M. *BioSci.* **62**, 962–969 (2012).
5. Toledo, V. M. & Barrera-Bassols, N. *La Memoria Biocultural* (Icaria, 2014).
6. Max, D. T. 'Green Is Good' *The New Yorker* (12 May 2014).
7. UNESCO Institute for Statistics. *Women in Science: UIS Fact Sheet* (UNESCO, 2012).
8. Carson, R. *Silent Spring* (Houghton Mifflin, 1962).
9. Alagona, P. S. *After the Grizzly: Endangered Species and the Politics of Place in California* (Univ. California Press, 2013).
10. Chapin III, F. S. et al. *Ecosphere* **2**, 89 (2011).

For a full list of co-signatories and further reading on this topic, see [go.nature.com/teztv](http://go.nature.com/teztv).

# A to-do list for the world's parks

Experts share their priorities for what must be done to make protected areas more effective at conserving global biodiversity.

**BOB PRESSEY**

## Maximize returns on conservation

Professor, Australian Research Council Centre of Excellence for Coral Reef Studies, James Cook University

Protected areas are meant to preserve biodiversity, but practice, measures of progress and targets do not reflect this role.

Governments and non-governmental organizations usually concentrate on politically palatable measures, such as numbers of hectares. Measures of progress and targets for protected areas should focus on placing protection where it can make the most difference.

A 2008 study estimated that only 7% of protected forests in Costa Rica would have been lost if not protected (K. S. Andam et al. *Proc. Natl Acad. Sci. USA* **105**, 16089–16094; 2008). These forests, like most protected areas worldwide, are in 'residual areas' — those where direct human threats to biodiversity are low, and where 'protection' makes

little difference. Misleadingly, target 11 of the Convention on Biological Diversity measures progress in percentages of land and sea protected. Meanwhile, the biodiversity of contested places continues to be eroded.

Performance metrics for protected areas should borrow from those in medicine, education and development. These fields all aim to maximize returns on investment. The language of programme evaluators is framed in terms of efficacy: what is the actual outcome of an intervention, compared with the outcome expected from no intervention?

For protected areas, efficacy means





Areas surrounding Victoria Falls, on the Zambia–Zimbabwe border, are protected by national-park status.

avoiding the loss of species and maintaining the integrity of ecosystems. There are methods for estimating the losses that protection has prevented (to provide lessons) or could prevent (to set priorities). By these metrics, protected areas can be disappointing.

Success depends on which natural resources societies are willing to leave unexploited. The trends are not encouraging. Australia, for example, hosts this year's World Parks Congress, but most of its terrestrial and marine parks are residual, and the country's protected-area strategy has no quantitative targets for avoiding loss. The congress could make a real difference if it steers policies away from meaningless, counterproductive targets. Each year of delay means avoidable, irreversible loss of biodiversity.

## DOUGLAS J. MCCAULEY

### Mega-parks need greater oversight

*Assistant professor in the department of ecology, evolution and marine biology, University of California, Santa Barbara*

In September, US President Barack Obama created the world's largest marine protected area network by massively expanding the Pacific Remote Islands Marine National Monument (PRIMNM). Collectively, the

PRIMNM is more than five times the size of the United Kingdom. Its creation ups the ante in a conservation phenomenon without precedent on land or sea. The eight marine mega-parks (each more than 250,000 square kilometres) announced in the past five years have almost doubled the amount of protected area in the oceans.

Three actions must be taken to ensure that mega-parks do more good than harm for the world's seas. First, governments must recognize that conventional forms of monitoring for protected areas are not tenable in parks that are larger than some countries. To ensure that areas such as the PRIMNM do not become 'paper parks' — marked as protected on maps but exploited in reality — governments must explicitly fund the development and use of next-generation enforcement, such as satellite and drone-based patrols. Such tools are not cheap, but mega-parks will not function unless they are designated in budgets as well as on maps.

Second, policy-makers must enact regulations to manage highly mobile animals in the 96% of the ocean left unprotected. Many of the most at-risk species (including some turtle, shark and marine mammal species) are not fully protected, even in parks as big as the PRIMNM.

Lastly, the marine mega-park movement does not let us off the hook for protecting crucial marine habitats at smaller scales. Bigger is better with marine protected areas, but these benefits might not scale linearly. Although establishing 100 strategically

placed, 10,000-square-kilometre marine parks is politically intractable, it would probably have done more for marine biodiversity than the establishment of just the PRIMNM.

If ineffectual practices can be avoided, environmental leaders will undoubtedly look back on this marine-mega-park era as one of the most important periods in the history of ocean conservation. If not, mega-parks will be little more than mega-hype.

## LANCE MORGAN

### Protect diverse marine habitats

*President of the Marine Conservation Institute*

A portfolio of well-protected, representative marine ecosystems — humankind's *in situ* seed vault for ocean life — is needed for biological and human resilience. Only about 2% of the ocean has any protection, and just 0.83% is 'no-take' reserves, where humans are not allowed to extract fish, oil or other resources. Marine biologists recommend that 20–30% of the ocean must be protected to maintain its biodiversity. This amount will provide enough abundance to restore depleted populations outside reserves.

To accelerate establishment of highly effective biodiversity refuges, the Marine Conservation Institute has initiated the



Global Ocean Refuge System (GLORES). The prestige and social capital that comes from receiving the GLORES status can spur governments, much as 'green building' certification has helped the adoption of sustainable practices in construction.

Capturing a diversity of habitats is key. US national parks, for example, often encompass mountainous areas of the United States, but not prairies and wetlands. GLORES considers the effect of a protected area in the context of others. It accounts for marine biogeography and connectivity; for example, kelp forests occur in temperate biogeographic regions, whereas coral reefs occur in tropical regions.

GLORES criteria require effective monitoring and enforcement, whether by communities, scientists or other authorities. The goal is to create protected areas in all of the different ocean regions and habitats (shallow and deep, sandy and rocky bottoms, and more).

GLORES will be easier, cheaper and faster to implement than many other approaches. Protecting places is much less knowledge-intensive and less costly than managing marine species one by one or persuading countries to protect areas one by one (often small areas that fishers care least about).

## HUGH POSSINGHAM Represent ecosystems

*Professor of mathematics and ecology, University of Queensland; Chair of conservation decisions, Imperial College London*

The Convention on Biological Diversity asks countries to conserve at least 17% of their land and 10% of their seas. It also calls for "ecological representation", the equitable coverage of species and habitats, but sets no quantitative targets. Representation is often ignored in designing systems of protected areas. For example, the koala is just one of many species that prefers under-protected fertile, well-watered habitats that are also favoured for agriculture and other development.

There are better approaches. Representation can be highly efficient. In 2004, the Great Barrier Reef Marine Park Authority used extensive economic and ecological data to create a system of 'no-take' protected areas that conserved at least 20% of every habitat while covering only 33% of the region.

To help refocus priorities, our group created software called Marxan. It uses mathematical optimization to prioritize places to design efficient and representative protected areas. We have also developed a



An anti-poaching team of the conservation group WWF on patrol in Minkébé National Park, Gabon.

new metric, protection equality, to measure equitable representation of habitats in a single number. It is a modification of the Gini coefficient commonly used to assess income inequality.

For example, the United States has a relatively large fraction of its land conserved, but its land protection equality is poor, only 0.33. Australia, which has a policy for representation, has a smaller fraction of its land conserved but a higher land protection equality of 0.51, a much more representative system. The protection equality of the Great Barrier Reef Marine Park is 0.80, more than any country. By contrast, the proposed re-zoning of Australia's Commonwealth waters is biased towards deeper waters and misses entire ecosystems.

We hope that this year's World Parks Congress will stimulate more-sophisticated tools for building representative systems of protected areas, and metrics for assessing them.

## LEE WHITE Manage parks professionally

*Executive secretary, Gabon National Parks Agency*

The special parts of our planet warrant and need exceptional stewardship. They are not getting it. We need a pact — between political leaders, civil society and conservation professionals — to increase the political capital of the environment.

Many of the world's rarest and most

iconic species — gorillas, chimpanzees, orangutans, elephants, lions, tigers and pandas — survive almost exclusively in protected areas. In both developed and developing countries, protected areas often contain the richest, most pristine ecosystems. They also provide crucial ecosystem services. Mangrove parks succour fisheries and protect against floods (see page 32); forests provide clean, reliable water and help to regulate the climate. Tourism and leisure use of parks improve people's quality of life. If preserved, the biodiversity within parks could well yield as-yet-unknown medicines and other products.

We need to strengthen and professionalize park management. Too many of the developing world's protected areas are chronically underfunded. And government neglect often means that management falls to international non-governmental organizations (NGOs). Because this responsibility is rarely formalized, NGOs do not have a strong mandate to protect these areas well.

For example, in the late 1990s, the government of Gabon failed to take responsibility for Minkébé National Park, so the conservation group WWF stepped in. Despite spending millions of dollars, WWF was unable to stop the slaughter of elephants there: at least 16,600 elephants were lost between 2004 and 2012, mainly to cross-border poachers. In May 2011, the government deployed 120 military personnel to support parks staff; in October 2014, with the situation still not under control, it pledged to double those numbers.

The people who fight to preserve our natural and cultural treasures must be trained and backed by their nations. Only then

ROBERT J. ROSS/GETTY





Coral gardens in the Palmyra Atoll and Kingman Reef national wildlife refuges, protected under the Pacific Remote Islands Marine National Monument.

will we be able to resist the ever-growing pressures that transnational crime, corruption and increasing population place on wildlife and wild lands.

## EMILY DARLING

### Conserve climate refuges

*David H. Smith conservation research fellow, University of North Carolina and the Wildlife Conservation Society*

Climate change raises a triple threat that existing marine protected areas were not designed to defend against. Warming, rising and acidifying seas threaten global marine biodiversity and ecosystem services. Even in protected areas, El Niño events and ocean heatwaves can bleach and destroy vast areas of healthy coral reefs — the canary in the coal mine of climate change.

To give coral reefs and other global ecosystems time to adapt, we need to identify areas that will escape the worst impacts of a changing climate. These should be protected as 'climate refuges' — areas that will experience less change over the coming decades. In the northern Mozambique channel and the Raja Ampat archipelago in Indonesia, for instance, upwelling and ocean gyres bring cool water that has allowed fragile corals to escape bleaching. Emerging evidence suggests that several million years ago, rare reef habitats that escaped rising temperatures provided the blueprint for contemporary diversity. Today, climate refuges may be

our best hope for protected areas to sustain healthy coral reefs into the future.

The first steps are to catalyse local communities, national governments and multilateral agencies to protect such areas. Urgently, we need to coordinate, fund and implement a global plan to link networks of climate refuges for all ecosystems: coral reefs, tropical rainforests, Arctic tundra and beyond. The World Parks Congress must lay the groundwork to incorporate climate refuges into conservation portfolios and protected areas.

## PETER J. S. JONES

### Assess governance structures

*Researcher on natural resource governance approaches, University College London*

Projects such as the International Union for Conservation of Nature's Green List are beginning to evaluate the effectiveness of protected areas systematically. This will help to shift the focus of conservation efforts from targets assessed just by hectares to other, more-meaningful objectives, focused on effectiveness. But to learn from successes and failures, we must also evaluate governance systems.

These systems incorporate roughly five approaches: top-down regulation, bottom-up participation, market mechanisms, awareness-raising and knowledge-sharing. We need to know what makes each effective,

and how these different approaches can be combined to reinforce each other.

On Chumbe Island, a private island park off the coast of Zanzibar in East Africa, diverse approaches mesh to form a strong governance framework. In 1994, a non-profit company was granted property rights to the island and its surrounding waters, along with obligations to the local environment and community. For example, income from ecotourists is invested in local schools and other community projects. The local police assist in enforcing a no-fishing zone, and anti-poaching patrols provide community services, such as helping fishing boats in peril.

Projects with fewer approaches are less robust. Consider the Cres-Lošinj Special Marine Reserve in Croatia. Here local authorities instituted the Adriatic Sea's largest marine protected area for dolphins, only for the designation to lapse when commercial developers touted the jobs and other economic benefits that a recreational marina could provide. In this case, top-down regulation, along with other governance approaches, might yet prove effective. The European Commission could oblige the Croatian government to reinstate protection as a condition of joining the European Union.

Too often, conservation discussions descend into unproductive debates about which governance approach is best, but the best solution varies with context. We need to learn the principles to match combinations of approaches with situations. The key to resilience is diversity — both of species in ecosystems and approaches in governance systems. ■





This former mangrove forest once provided benefits such as storm protection and carbon sequestration.

# Account for depreciation of natural capital

Economic indicators that omit the depletion and degradation of natural resources and ecosystems are misleading, warns **Edward B. Barbier**.

For the past year, academics and policy-makers have been discussing Thomas Piketty's 2013 economics best-seller, *Capital in the Twenty-First Century*<sup>1</sup>. It documents the considerable rise over the past 40 years in national wealth relative to national income in eight of the richest economies — the United States, Japan, Germany, France, the United Kingdom, Italy, Canada and Australia. The national wealth of each of these countries increased from 2–3 times national income in 1970 to 4–6 times income in 2010.

Piketty relies on standard income conventions as prescribed in the United Nations national accounts. He includes natural resources such as fossil fuels, minerals and forests in his estimate of a country's capital. But his measures of national income and savings adjust only for depreciation of 'fixed capital' — buildings, equipment and so on.

We must also account for the depreciation of natural capital in appraising wealth. This is the value of net losses to natural resources, such as minerals, fossil fuels, forests and

similar sources of material and energy inputs into our economy. If we use up more natural capital to produce economic output today, then we have less for production tomorrow.

At the same time, we are also squandering valuable ecological capital — ecosystems provide important goods and services to the economy, such as recreation, flood protection, nutrient uptake, erosion control, water purification and carbon sequestration. By converting and degrading ecosystems, we are depreciating this important ecological capital endowment.

Economic indicators change dramatically when the depletion and degradation of natural resources and ecosystems are accounted for. Here, I show by how much, through a worked example of mangroves in Thailand. Depreciation of natural capital is particularly high in developing economies, which are often rich in resources and ecosystems. We must retool our measures of income and wealth accordingly, starting with net domestic product.

## CREATIVE ACCOUNTING

Since 1970, the World Bank's World Development Indicators have provided estimates for most countries of the adjustments to national income, income growth and savings that arise from net depletion of forests, energy resources and minerals. This rate of natural-capital depreciation as a percentage of adjusted net national income over the past four decades is alarming (see 'Natural capital').

Two global trends are noticeable. First, the decline in natural capital has been five times greater on average in developing economies than in the eight richest countries. Second, natural capital depreciation in all countries has risen significantly since the 1990s. There was a dip during the global recession of 2008–09, but as the world economy has recovered, so has the rate of resource use.

Ecological capital, too, is clearly endangered by current patterns of economic development. Over the past 50 years, ecosystems have been modified more rapidly and extensively than in any comparable period in human history, largely to meet burgeoning demands for food, fresh water, timber, fibre and fuel. According to the worldwide Millennium Ecosystem Assessment, approximately 60% of major global ecosystem services have been degraded or used unsustainably, including fresh water, wild fisheries, air and water purification, and the regulation of regional and local climate, natural hazards and pests.

Unfortunately, ecological capital, being unique, poorly understood and difficult to measure, tends to be undervalued. Consider the example of mangroves in Thailand from 1970 to 2009<sup>2</sup>. Average annual mangrove loss in Thailand has fallen steadily in every decade since the 1970s. Yet cumulatively,



Thailand is estimated to have lost around one-third of its mangroves since the 1970s, mainly to the expansion of shrimp farming and other coastal development.

### MANGROVE ECONOMICS

Mangroves provide four essential ecosystem benefits: wood and products such as shellfish, plants, honey and medicines; nursery and breeding grounds for offshore fisheries; storm protection; and carbon sequestration.

I use estimates of these benefits to determine the annual net gain or loss in mangrove value resulting from conversion to other land uses. This net value has two components. The remaining mangroves generate extra benefits each year that do not appear in the national accounts, such as net subsistence for local coastal communities and economy-wide carbon-sequestration benefits. From these values, I subtract the net loss in land value that arises each year from converting mangroves to some other economic activity, such as shrimp farming.

The economic impacts are significant. During the 1970s and 1980s, when mangrove deforestation was rapid, Thailand lost US\$1.69 and \$0.76, respectively in mangrove net values per person per year. By 2009, around one-third of the 1970 mangrove area was deforested and Thailand's population had grown rapidly. As a result, the total value from the subsistence and carbon benefits of the remaining mangroves has halved, from \$0.57 to \$0.28 per person per year (see 'Cutting costs'). This means that even though mangrove loss slowed in the 1990s and 2000s, the net values of mangroves were very modest, only \$0.11 and \$0.25 respectively.

To put it another way, cumulative mangrove deforestation over the past four decades in Thailand has cost each Thai citizen \$40. This debit amounts to losses of more than \$2.73 billion, which have never appeared in Thailand's national accounts.

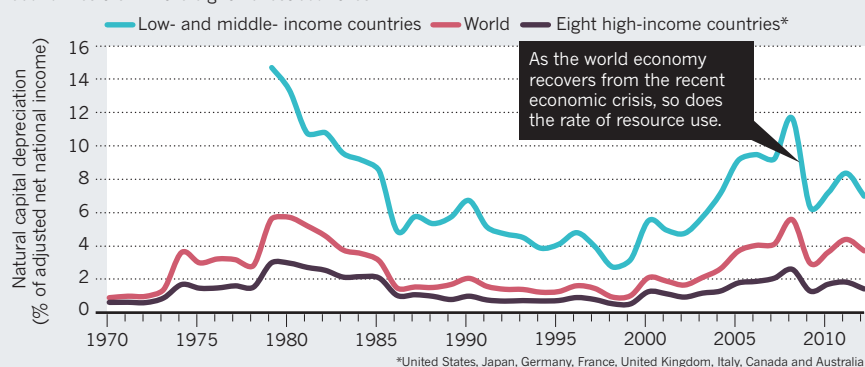
### COUNTING THE COST

Many more examples are now needed — for different countries and regions, and for other key ecosystems, such as tropical forests, coral reefs, freshwater wetlands, grasslands and so on.

There are three caveats. First, there are clearly intrinsic values to preserving unique natural resources, species and ecosystems, as well as the biological diversity contained in these systems, which are not captured by such an approach. Second, the benefits of many important ecosystem services are difficult to value, such as pollution control, pollination, climate regulation and watershed protection. Third, measures of natural-resource depletion need to move beyond minerals, energy and timber harvests to include other vital resources, such as soils, air quality, aquifers,

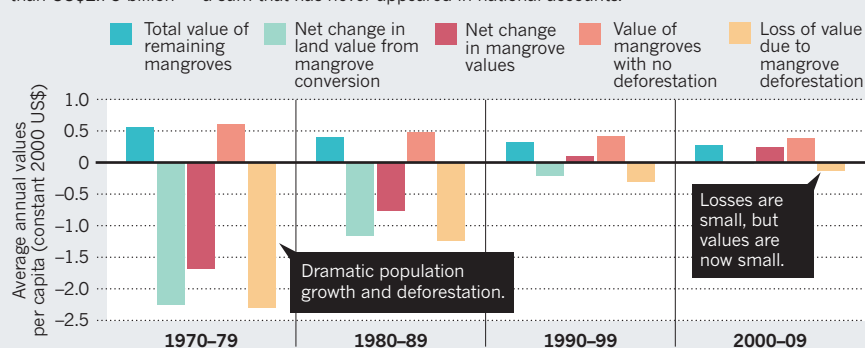
### NATURAL CAPITAL

The decline in natural capital has been five times greater on average in developing economies than in the eight richest countries.



### CUTTING COSTS

Losing one-third of its mangroves to deforestation since 1970 has cost Thailand more than US\$2.73 billion — a sum that has never appeared in national accounts.



fisheries and non-timber forest resources.

The UN and the World Bank have begun pilot studies to construct adjustments to income and wealth that include changes in ecological capital. The UN *Inclusive Wealth Report 2012* has developed<sup>3</sup> accounts from 1990 to 2008 for 20 countries that include non-timber benefits from forests, carbon sequestration, fisheries (for four countries only), carbon damages and agricultural land, as well as minerals, energy and timber. The World Bank is expanding pilot studies on ecosystem accounting from 8 to 15 developing countries, which cover water, forest and mangrove ecosystems (see [www.wavespartnership.org](http://www.wavespartnership.org)).

For estuarine and coastal ecosystems, there are already 80 valuation estimates from all over the world for storm protection, erosion control, water purification and supply, carbon sequestration, recreation and maintenance of fishing, hunting and foraging activities — and the list is growing<sup>4</sup>.

What will it take to move beyond these encouraging pilot studies? The UN systems of national accounts must adopt a more systematic approach that all countries can follow to account for losses of natural capital and ecological capital, as we already do for fixed capital depreciation. And, in the case of complex ecosystems and landscapes, we need to resolve problems of 'double

counting' ecosystem services that might serve as 'inputs' into production or that are provided by multiple ecosystems, such as the protection of shorelines simultaneously by coral reefs, seagrass beds and mangroves.

Piketty might be right that, since 1970, there has been substantial accumulation of capital relative to income in the rich countries of the world. As low- and middle-income countries try to emulate this success, they will also be striving to accumulate more wealth. But as my estimates show, our economies have been trading one form of capital, Earth's riches, for another — human riches. Without accounting accurately for this trade-off, we will continue to have a false impression of economic progress and growth. That is as dangerous as flying an aeroplane into the night without navigation tools or instruments. ■

**Edward B. Barbier** is professor of economics at the University of Wyoming, Laramie, Wyoming, USA.  
e-mail: [ebarbier@uwyo.edu](mailto:ebarbier@uwyo.edu)

1. Piketty, T. *Capital in the Twenty-First Century* (Harvard Univ. Press, 2013).
2. Barbier, E. B. *Environ. Dev. Econ.* **18**, 133–161 (2013).
3. UNU-IHDP and UNEP *Inclusive Wealth Report 2012: Measuring Progress Toward Sustainability* (Cambridge Univ. Press, 2012).
4. Barbier, E. B. *Resources* **2**, 213–230 (2013).

## ACCELERATOR PHYSICS

# Surf's up at SLAC

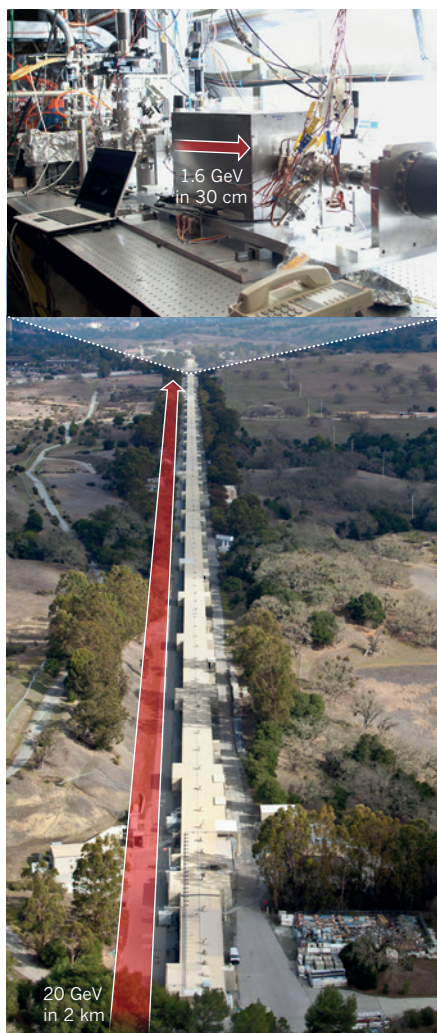
A 'plasma afterburner' just 30 centimetres long accelerates electrons hundreds of times faster than giant conventional accelerators. The result may ultimately open up a low-cost technology for particle colliders. [SEE LETTER P.92](#)

MIKE DOWNER & RAFAL ZGADZAJ

In November 2012, Guinness World Records reported that 120 surfers in Australia rode the same wave simultaneously for more than 5 seconds<sup>1</sup>. "The trick was to get them all to do the same thing at the same time," said group leader Wes Smith. "It was an operation of military-like precision and we finally got there." Now Litos and colleagues, in work at the SLAC National Accelerator Laboratory in Menlo Park, California, reported on page 92 of this issue<sup>2</sup>, have 'got there', too, by surfing half a billion 20-billion-electronvolt electrons on a steep charge-density wave about the size of a marine phytoplankton, travelling through ionized gas (plasma). The wave was driven by a companion electron bunch as it raced at nearly the speed of light through a 30-centimetre-long chamber filled with plasma (Fig. 1).

Although this inaugural experiment lost about 90% of its 'surfers' along the way, the surviving electrons gained 1.6 billion electronvolts, or 1.6 gigaelectronvolts (GeV), in energy with unparalleled uniformity, maintaining roughly 1% energy spread throughout their wild ride, while sucking away an unprecedented fraction (up to 30%) of the wave's energy. Such uniform, efficient acceleration required the researchers to inject the surfing electron bunch into the wave, and to adjust the bunch's charge and shape, with a military-like precision made possible by SLAC's recently commissioned US\$15-million Facility for Accelerator Science and Experimental Tests (FACET)<sup>3</sup>. Because the plasma wave accelerated electrons 500 times faster than SLAC's main particle accelerator, the result might herald a new generation of compact 'plasma afterburners' that could boost the energy of conventional particle accelerators and potentially reduce the skyrocketing cost of high-energy physics machinery<sup>4</sup>.

Seven years ago, before FACET was even proposed, the same team had used single bunches of about 10 billion 42-GeV electrons and accelerated them over the full 3.2-kilometre length of SLAC's main machine to drive a similar plasma wave<sup>5</sup>. A handful of electrons from the tail of the drive bunch were caught in the drive bunch's wake and were accelerated up to 84 GeV, twice the energy of the electrons in the original drive bunch, within a metre-long



**Figure 1 | Ramping up the energy.** SLAC's main accelerator, shown in aerial view, accelerates electron bunches from 0 to 20 GeV energy over 2 km, which amounts to adding 0.01 GeV to each electron every metre. The new Facility for Accelerator Science and Experimental Tests (FACET) used by Litos *et al.*<sup>2</sup> then splits each 20-GeV bunch into two independently controlled tandem bunches. The leading bunch creates a new micro-accelerator inside a 30-cm chamber (top), in which it drives a charge-density wave in ionized gas, much as a boat drives a wake in water. The trailing bunch rides the lead bunch's wake and, when optimally positioned, extracts up to 30% of its energy, boosting each electron's energy by 1.6 GeV in only 30 cm.

plasma chamber. However, the electrons emerging from this first-generation plasma afterburner ranged in energy from less than about 35 GeV to 84 GeV, more electrons were decelerated than accelerated, and most of the energy of the plasma wave was left untapped. FACET — which now shares SLAC with the Linac Coherent Light Source, and thus starts with 20-GeV electrons accelerated over part of SLAC's length — was designed to correct these shortcomings. The facility exploits new particle-beam technology to split the SLAC bunches into two tandem bunches whose time separation, charge and shape are, with some limits, independently controllable.

In the new experiments, the researchers used a little over half of the 20-GeV SLAC bunch to drive a plasma wave, and then timed its nearly equally charged twin to surf just a hair's breadth behind, where its core rode the enormous electrostatic field of the drive bunch's wake. Without the trailing surfing bunch, this field would be far from uniform, varying from 3 billion to 10 billion volts per metre (fields stronger than ordinary, non-plasma matter can withstand) just over the tiny region in which the surfing bunch was so painstakingly positioned.

Had the researchers injected a lower-charged surfing bunch, it would have suffered the same fate as in the earlier experiment by broadening in energy. This would render it useless for high-energy physics applications, which require particle energy to be tuned precisely to create and identify new particles, such as the Higgs boson. However, Litos *et al.* took advantage of physics learned from computer simulations<sup>6</sup> showing that a high-charge surfing bunch could 'load' the plasma wake, flattening its electrostatic fields locally. It is as if the 120 Australian surfers had sufficient collective weight to flatten the curved ocean wave into an inclined plane so that they could all accelerate at the same rate. This trick solved two problems simultaneously: it enabled a high-charge bunch to accelerate nearly monoenergetically while maximizing energy extraction from the plasma wake.

Can plasma surfing meet future needs of high-energy physics research, which include electron bunches with sufficiently high energy, charge, repetition rate and focusability that they

SLAC NATL ACCELERATOR LAB



can create detectable amounts of new particles that may be lurking in the cosmic underworld? The jury is still out. The present 1.6-GeV energy gain (starting from 20 GeV) is no greater than that achieved by plasma accelerators driven by light pulses from lasers (starting from zero)<sup>7</sup>, which are much smaller and less-expensive instruments than SLAC. Nevertheless, electron-driven plasma accelerators scale more readily to gains of tens of gigaelectronvolts than do their laser-driven counterparts, as demonstrated in previous work<sup>5</sup>.

Improved bunch-shaping technology will better match surfing bunch to plasma wave, increasing electron-survival rate and thus the number of accelerated electrons. Yet the Higgs boson has a mass equivalent to 126 GeV, and physical theories such as supersymmetry predict additional particles that have even greater mass than the Higgs and may be the source of the elusive 'dark matter' that seems to comprise about 25% of the Universe. Creating and identifying these new denizens of the Universe could set the next energy frontier at many thousands of gigaelectronvolts. Reaching these energies will probably require synchronized, multi-staged plasma accelerators — a daunting, and largely unexplored, technical challenge in view of the micrometre dimensions of plasma waves.

An interesting alternative proposal is to drive plasma waves with very energetic proton bunches, which because of their greater mass can push plasma waves for hundreds of metres, potentially accelerating electrons to the energy frontier in a single stage<sup>8</sup>. In either case, plasma acceleration of positrons (anti-electrons) lags far behind electron acceleration because plasma waves shaped like those in the current experiment defocus surfing positron bunches, degrading their usefulness. Positron acceleration is important because high-energy collisions of electrons and positrons, a natural matter-antimatter pair, create a richer collection of products with higher efficiency than, say, electron-electron collisions, and thus offer one of the most promising routes to particle discovery. FACET, with access to SLAC's companion positron beam, is uniquely positioned to explore new ways to shape plasma waves in order to advance plasma-based positron acceleration.

Finally, even if the energies and charges required for an electron-positron collider are achieved, debate rages over whether focused, plasma-surfed particle beams can yield particle-discovery events at rates competitive with those achieved with conventional accelerator technology<sup>9–11</sup>, which underlies proposed tens-of-kilometres-long machines such as the International Linear Collider and the Compact Linear Collider. These uncertainties notwithstanding, Litos *et al.* have overcome one of the most difficult challenges so far in the long quest for small, affordable accelerators, and have given the plasma-surfing community every reason to surge ahead. ■

**Mike Downer and Rafal Zgadzaj** are in the *Physics Department, University of Texas at Austin, Texas 78712-1081, USA.*  
e-mail: [downer@physics.utexas.edu](mailto:downer@physics.utexas.edu)

1. [www.worldrecordacademy.com/sports/most\\_surfers\\_riding\\_the\\_same\\_wave\\_120\\_surfers\\_set\\_world\\_record\\_113137.html](http://www.worldrecordacademy.com/sports/most_surfers_riding_the_same_wave_120_surfers_set_world_record_113137.html)
2. Litos, M. *et al.* *Nature* **515**, 92–95 (2014).
3. Hogan, M. J. *et al.* *New J. Phys.* **12**, 055030 (2010).
4. Lee, S. *et al.* *Phys. Rev. ST Accel. Beams* **5**, 011001 (2002).

5. Blumenfeld, I. *et al.* *Nature* **445**, 741–744 (2007).
6. Tzoufras, M. *et al.* *Phys. Rev. Lett.* **101**, 145002 (2008).
7. Wang, X. *et al.* *Nature Commun.* **4**, 1988; <http://dx.doi.org/10.1038/ncomms2988> (2013).
8. Caldwell, A., Lotov, K., Pukhov, A. & Simon, F. *Nature Phys.* **5**, 363–367 (2009).
9. Schroeder, C. B., Esarey, E. & Leemans, W. P. *Phys. Rev. ST Accel. Beams* **15**, 051301 (2012).
10. Lebedev, V. & Nagaitsev, S. *Phys. Rev. ST Accel. Beams* **16**, 108001 (2013).
11. Schroeder, C. B., Esarey, E. & Leemans, W. P. *Phys. Rev. ST Accel. Beams* **16**, 108002 (2013).

## DEVELOPMENTAL BIOLOGY

# Cells unite by trapping a signal

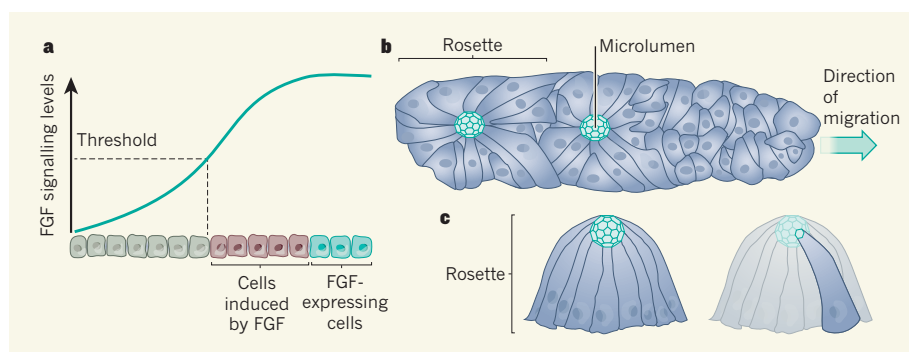
**Gradients of fibroblast growth factors often induce cells to adopt different fates. A study in zebrafish embryos reveals another, unexpected role when the factors are trapped in small spaces by a special arrangement of cells. SEE LETTER P120**

JAMES SHARPE

**B**uilding complex, multicellular organs during embryonic development is not just about making different cell types, it is about getting the right cells in the right place. For a cell to have some sense of where it is, it must integrate diffusible signals released by its neighbours. On page 120 of this issue, Durdu *et al.*<sup>1</sup> provide evidence for a surprising new way by which diffusible signals such as fibroblast growth factors (FGFs) are controlled — by trapping them in small, closed extracellular spaces called microlumina, from which they have access to only a discrete collection of cells.

Exactly how signalling molecules provide

enough spatial information to build complex organisms is still obscure, but studies of the principles of signalling generally split into two types. Those of the upstream part of signalling ask how the movement of diffusible molecules — sometimes called morphogens — is controlled to form appropriate spatial gradients<sup>2,3</sup>, for example by 'sticky' molecules in the extracellular matrix<sup>4</sup>. Studies of the downstream part ask how these spatial distributions are used by receiving cells to control cellular 'decisions'. A well-characterized example of the upstream part is the FGF family of secreted proteins<sup>5</sup>. These often form coherent spatial gradients within which different levels of signalling divide the responding cell



**Figure 1 | Roles of fibroblast growth factors (FGFs).** **a**, A common role for FGFs is to induce different cell fates through spatial variations in FGF levels determined by the distance from FGF-expressing cells. In this schematic, cells that experience FGF levels above a threshold value are induced, whereas the others are unaffected. **b**, During the development of zebrafish embryos, a group of cells called the primordium, shown here from above, migrates from near the head-end to the tail. As it migrates, cells cluster into rosette structures that drop off at regular intervals and then develop into mechanosensory organs. Durdu *et al.*<sup>1</sup> report that confinement of FGFs to the microlumen at the centre of a rosette coordinates the cells within that cluster, so that the rosette drops off in a well-organized manner. **c**, The microlumina are enclosed by a patchwork of membrane sections contributed by all the surrounding cells, as shown in this side view of a rosette. The right-hand graphic shows how one cell contributes to the microlumen, which is not shown to scale.

population into sub-groups with different fates (Fig. 1a).

Durdu and colleagues took a fresh look at FGFs in their study of the development of the zebrafish lateral line — a sensory organ that lies along either side of all fishes, allowing them to sense vibrations in the water. In this developmental process, about 100 cells (called the lateral-line primordium) start near the head-end of the embryo and, over a two-day period, collectively migrate along the entire length of the developing body under the skin towards the tail<sup>6</sup>. During this journey, subgroups of cells cluster together within the primordium. These are called rosettes, because the cells adopt a radial arrangement in which each cell has an extension towards an apparent central common connection point (Fig. 1b). As the primordium migrates along the body, it drops off these rosettes one by one at regular intervals. Each rosette goes on to develop into a discrete mechanosensory organ.

The authors knew that manipulating FGFs can affect the spacing of dropped organs<sup>7</sup>, but not whether this was through a general effect on primordium velocity. They therefore quantified time-lapse movies of developing zebrafish embryos in which Fgf3 levels had either been upregulated by overexpression or repressed by drug inhibition. In both cases, they saw that the migratory velocity of the primordium was unaltered, which means that Fgf3 was affecting the drop-off frequency instead.

Having established a clear link between FGF signalling and rosette drop-off, Durdu *et al.* next explored where the signalling occurs. Fluorescence imaging of Fgf3 attached to green fluorescent protein suggested that it was localized into small, concentrated volumes at the apical centre of each rosette. Correlative microscopy (which combines fluorescence microscopy with electron microscopy) then revealed a striking cell-membrane arrangement: at the apical centre of each rosette was a microlumen formed by the cell membranes of all the cells of that rosette (Fig. 1c).

The researchers again used time-lapse imaging to show that the moment when Fgf3 starts to accumulate in a microlumen correlates with the time when that rosette begins to slow down in preparation for dropping out of the primordium. This pointed towards the intriguing possibility that FGF signalling is used on a very local basis to control the behaviour of just the 20 or so cells of one rosette. Durdu and co-workers went on to use all the advantages of the zebrafish system — ease of genetic modification and micromanipulation, and its suitability for high-quality time-lapse imaging — to test the idea.

They modified a single rosette so that one of its cells had increased Fgf3 levels (using either single-cell transplantation or a stochastic inducible genetic system), and observed that just this rosette was forced to drop out early

from the primordium. On average, neither the rosettes before nor after it were prematurely dropped. To perform the opposite experiment, they punctured microlumina with a laser, thereby letting Fgf3 leak out. Satisfyingly, they observed the expected delay in rosette drop-off, again without affecting the previous or subsequent rosettes.

Several questions are not addressed in the study: for example, how the microlumina form in the first place; how levels of FGF expression are controlled; and, perhaps most directly relevant to the authors' findings, how FGF signalling accelerates rosette drop-off. But the strength of Durdu and colleagues' experiments is that single rosettes were manipulated *in vivo*, thus providing evidence that the microlumen can indeed restrict FGF signalling to the cells of just one rosette.

In this system, FGFs do not adopt one of their conventional upstream roles, in which a coherent swathe of different signalling levels splits a responding population of cells. Instead, the microlumen forces FGFs to take on a more downstream role: coordinating the response to a morphogenetic event, and ensuring that all cells of the rosette respond while none of the neighbours do. It is an intriguing case of multicellular architecture feeding back to control molecular signalling directly.

Because FGF concentrations accumulate only when the microlumen is topologically complete, the factors also provide a temporal checkpoint to the process. It thus unites a group of cells both temporally and spatially in a coordinated all-or-nothing response. This is an interesting, and slightly surprising, way to use a highly diffusible signalling molecule, but may turn out to be a widely employed mechanism in nature. ■

**James Sharpe** is in the Systems Biology Program, Centre for Genomic Regulation, 08003 Barcelona, Spain; at the Universitat Pompeu Fabra, Barcelona; and at the Institució Catalana de Recerca i Estudis Avançats, Barcelona.  
e-mail: james.sharpe@crg.eu

1. Durdu, S. *et al.* *Nature* **515**, 120–124 (2014).
2. Dubrulle, J. & Pourquié, O. *Nature* **427**, 419–422 (2003).
3. Lander, A. *Cell* **128**, 245–256 (2007).
4. Häcker, U., Nybakken, K. & Perrimon, N. *Nature Rev. Mol. Cell Biol.* **6**, 530–541 (2005).
5. Dorey, K. & Amaya, E. *Development* **137**, 3731–3742 (2010).
6. Ghysen, A. & Dambly-Chaudière, C. *Genes Dev.* **21**, 2118–2130 (2007).
7. Nechiporuk, A. & Raible, D. W. *Science* **320**, 1774–1777 (2008).

This article was published online on 22 October 2014.

#### ASTROPHYSICS

## Monster star found hiding in plain sight

**Massive stars are rare, but they are sources of some of the most energetic phenomena seen in the Universe today. A high-mass candidate has now been found in a star-forming region that has been observed for more than 50 years.**

**DONALD F. FIGER**

The most massive stars in the Universe captivate the imagination of laymen and experts alike. They represent an extreme form of star and produce outsized effects on their environment. Although stars with masses greater than 20 times the Sun's mass comprise only about 1% of all stars in a young star cluster, their ionizing radiation, stellar winds and ejecta from supernovae dominate some of the most observable phenomena in the Galaxy. Massive stars are among the few bodies that can be seen in other galaxies, and they are probably linked to the most massive explosions in the Universe. Finally, they are thought to have seeded the early Universe with heavy elements (those heavier than helium), which are now seen in even the oldest stars. Writing in *Astronomy & Astrophysics*,

Wu *et al.*<sup>1</sup> identify the next heavyweight contender — a star with the decidedly unsexy name of W49nr1.

Wu and colleagues claim a mass for this star that would place it among the most massive known, but a sceptic might say “extraordinary claims require extraordinary evidence”. Indeed, astronomers have, on further inspection, often thrown such assertions on the rubbish heap of history.

This kind of claim relies on models that translate the amount of observed starlight into an estimate of the mass of the star. Generally, the more massive the star, the brighter it is. As is almost always the case, Wu *et al.* observe light from the star over only a fairly narrow range of wavelengths, representing much less than 1% of the total emitted light. It would be useless to convert that relatively small portion of the total light into a mass estimate were it



**Donald F. Figer** is at the Center for Detectors, Rochester Institute of Technology, Rochester, New York 14623-5603, USA.  
e-mail: figer@cfdr.rit.edu

1. Wu, S.-W. *et al. Astron. Astrophys.* **568**, L13 (2014).
2. Zhang, B. *et al. Astrophys. J.* **775**, 79 (2013).

3. Cassinelli, J. P., Mathis, J. S. & Savage, B. D. *Science* **212**, 1497–1501 (1981).
4. Feitzinger, J. V., Schlosser, W., Schmidt-Kaler, T. & Winkler, C. *Astron. Astrophys.* **84**, 50–59 (1980).
5. de Koter, A., Heap, S. R. & Hubeny, I. *Astrophys. J.* **509**, 879–896 (1998).
6. Crowther, P. A. *et al. Mon. Not. R. Astron. Soc.* **408**, 731–751 (2010).
7. Figer, D. F. *Nature* **434**, 192–194 (2005).

8. Chené, A.-N., Schnurr, O., Crowther, P. A., Lajus, E. F. & Moffat, A. F. J. *IAU Symp.* **272**, 497–498 (2010).
9. Conti, P. S. & Blum, R. D. *Astrophys. J.* **564**, 827–833 (2002).
10. Alves, J. & Homeier, N. *Astrophys. J.* **589**, L45–L49 (2003).
11. Westerhout, G. *Bull. Astron. Inst. Neth.* **14**, 215–260 (1958).

## ECOLOGY

# Diversity breeds complementarity

Evolutionary and ecosystem processes have long been treated as distinct. The finding that interactions among plant species cause rapid evolutionary changes that affect ecosystem function suggests that it is time for unification. [SEE LETTER P.108](#)

DAVID TILMAN & EMILIE C. SNELL-ROOD

The great naturalist Charles Darwin proposed his theory of evolution by natural selection as a unifying explanation for patterns seen in the natural world. But the unity sought by naturalists gave way to more-fragmented perspectives as natural history itself speciated into the modern disciplines of ecosystem ecology, community ecology, population biology, palaeontology and evolution. In this issue, Zuppingner-Dingley and collaborators<sup>1</sup> (page 108) have taken a significant step towards a reunification of these disciplines. Their findings in an experimental study of plants suggest that ecosystem and evolutionary processes cannot be separated: ecological interactions among a large number of plant species can cause rapid evolutionary changes that, in turn, influence ecosystem processes.

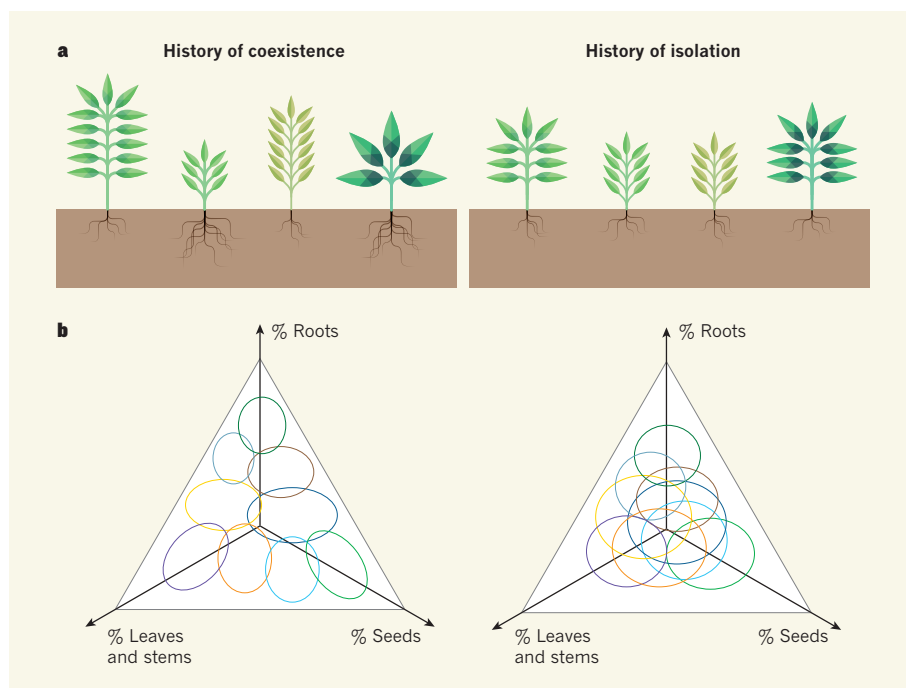
An idea central to both ecology and evolution is that of the niche — the set of environmental conditions in which a particular species thrives. In ecology, niche differences among species help to explain why large numbers of competing species coexist, and why greater plant diversity leads to greater ecosystem productivity<sup>2</sup>. In evolutionary biology, the niche concept features prominently in our understanding of how new species arise. Competition between closely related species drives the evolution of trait differentiation, such as bird beaks that are specialized for different seeds or lizard limbs that are suited for either climbing or walking. The evolution of such character displacement can be seen in laboratory experiments using microorganisms<sup>3</sup> and in field studies of incipient species formation, such as in Darwin's finches on the Galapagos Islands<sup>4</sup>.

In their study of character displacement, Zuppingner-Dingley and collaborators made use of experimental field plots in which

16 species of grassland plant were grown either in monocultures or in mixed plots of 4 or more species for 8 years. They then collected these plants, propagated them in the lab, and assembled the offspring in new communities: either monocultures or mixed communities of two species. They observed that, relative to the

monocultures, the 8-year period of selection in the high-diversity communities caused shifts in the traits of the plant species, specifically in plant height and leaf thickness. These shifts were consistent with character displacement and niche differentiation (Fig. 1a). The researchers also observed ecosystem-level consequences of these rapid evolutionary changes: the mixed cultures of plants from the diverse communities were more productive in terms of biomass than were mixed cultures from monocultures. These results exemplify the emerging field of eco-evolutionary dynamics, which emphasizes that not only does ecology drive evolution, but evolutionary change feeds back to affect ecological processes<sup>5</sup>.

In Zuppingner-Dingley and colleagues' study, laboratory propagation of the plants increased the chance that the differences between the high- and low-diversity selection groups were due to genetic divergence.



**Figure 1 | Evolutionary niche shifts.** **a**, Zuppingner-Dingley *et al.*<sup>1</sup> find that, when plant species are grown in a common environment, those that have a history of selection in diverse communities develop greater differences in traits (such as height and leaf thickness) than species that have a history of isolation. **b**, This idea feeds into our understanding of how evolutionary history influences the ecological interactions of species that compete for growth factors such as soil nutrients, light and space. All species face trade-offs. For instance, biomass that is allocated to obtaining soil nutrients (roots) cannot be used to obtain light (leaves and stems) or to disperse to open sites (seeds). Graphically depicted, the resulting 'trade-off surface' (triangles) represents all possible ways in which plant species (ellipses) can allocate their biomass. A history of selection in diverse communities results in greater interspecific differences (less overlap of ellipses) and more specialization (smaller ellipses) than a history of isolation.

## ORGANIC CHEMISTRY

# Shape control in reactions with light

The report of a light-activated catalyst that dictates the three-dimensional shape — the stereochemistry — of molecules formed in an organic reaction suggests a new strategy for controlling such reactions using visible light. [SEE LETTER P.100](#)

However, it is possible that epigenetic factors — heritable changes that do not involve DNA-sequence changes — could have had a simultaneous role<sup>6</sup>. If so, this invokes a broader question<sup>7</sup> concerning the influence of developmental plasticity in niche differentiation: might character displacement in a diverse community initially be driven by developmental responses to resources and competitors that are later genetically assimilated as speciation occurs? The present study suggests that this challenging question could be addressed in real time with an experimental, field-based approach.

Because natural communities are diverse, selective forces that emerge from interactions between many species may be unexpectedly influential factors that shape species traits. How might we conceptualize this possibility? Stable coexistence requires both character displacement and evolutionarily unavoidable trade-offs between species<sup>8,9</sup>. Such intraspecific trade-offs occur only if allocation of biomass to traits that increase performance in one type of environment decreases performance in other environments. For example, plants that have greater root mass perform better in infertile soils, but those that have more leaf and stem mass — and thus are taller and capture more light — dominate fertile soils. Such trade-off 'surfaces' could explain how the ecological interactions that allow multi-species coexistence also influence the rate and pattern of species formation<sup>8,9</sup> (Fig. 1b).

Although the disciplines of ecosystem ecology and evolution have developed their own perspectives, if each incorporated elements of the other, both disciplines would be strengthened. It is time for a reunification of all of the branches of natural history in a renewed search for unified explanations of the patterns seen in the natural world. ■

**David Tilman and Emilie Snell-Rood** are in the Department of Ecology, Evolution and Behavior, University of Minnesota, St Paul, Minnesota 55108, USA. D.T. is also in the Bren School of Environmental Science and Management, University of California, Santa Barbara, USA.  
e-mails: [tilman@umn.edu](mailto:tilman@umn.edu); [emilies@umn.edu](mailto:emilies@umn.edu)

1. Zupping-Dingley, D. *et al. Nature* **515**, 108–111 (2014).
2. Lehman, C. L. & Tilman, D. *Am. Nat.* **156**, 534–552 (2000).
3. Rainey, P. B. & Travisano, M. *Nature* **394**, 69–72 (1998).
4. Grant, P. R. & Grant, B. R. *Science* **313**, 224–226 (2006).
5. Fussmann, G. F., Loreau, M. & Abrams, P. A. *Funct. Ecol.* **21**, 465–477 (2007).
6. Verhoeven, K. J. F., Jansen, J. J., van Dijk, P. J. & Biere, A. *New Phytol.* **185**, 1108–1118 (2010).
7. Pfennig, D. W. *et al. Trends Ecol. Evol.* **25**, 459–467 (2010).
8. Tilman, D. *Proc. Natl Acad. Sci. USA* **101**, 10854–10861 (2004).
9. Tilman, D. *Am. Nat.* **178**, 355–371 (2011).

This article was published online on 15 October 2014.

KAZIMER L. SKUBI & TEHSHIK P. YOON

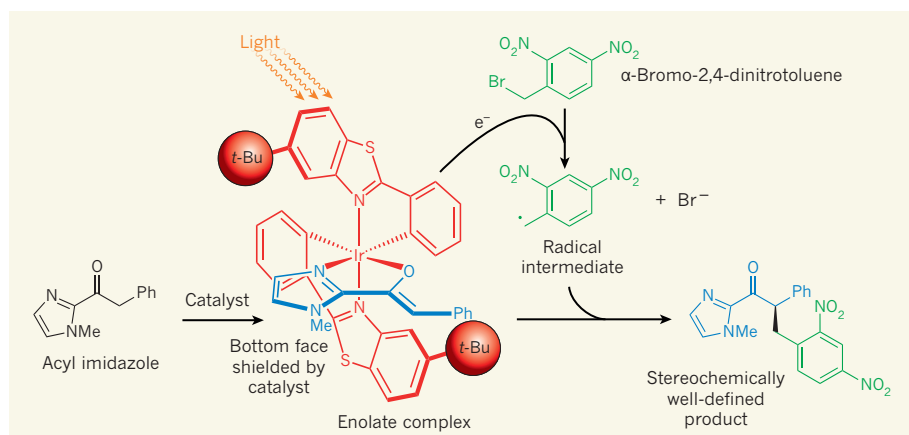
Photochemical reactions can occur when a molecule absorbs light. Such reactions are greatly valued by organic chemists for their ability to promote fascinating changes in molecular structure that cannot be replicated in any other way. However, the application of these reactions for syntheses has long been hindered by several practical limitations. One of the biggest is the dearth of effective strategies for controlling the three-dimensional shape of the organic molecules produced. On page 100 of this issue, Huo *et al.*<sup>1</sup> report an elegant approach to address this long-standing challenge.

The interaction between light and matter constitutes one of the most active areas of scientific research. This year, for instance, the Nobel prizes for physics and chemistry were awarded for the development of efficient light-emitting devices and for the use of fluorescence in ultra-high-resolution microscopy, respectively. From energy science to biomedicine to materials engineering, photochemistry is a vibrant theme

of such research, transecting many fields.

Photochemical reactions have been used to streamline complex syntheses and to build structurally unusual organic frameworks. However, organic molecules are generally transparent to visible light — they cannot absorb its energy for use in chemical reactions. Organic photochemical reactions have typically needed ultraviolet light, which requires specialized equipment and instrumentation capable of handling high-energy ultraviolet photons. This has limited the study of photochemical synthesis to a fairly small community of specialists.

But in the past several years, a variety of exciting photoreactions have been developed that use visible light, and so can be carried out with simple household light sources or even sunlight<sup>2</sup>. The key insight was that certain transition-metal complexes (typically based on ruthenium or iridium) that absorb relatively low-energy wavelengths of visible light can be used as catalysts to activate a wide range of organic substrates, thereby enabling new reactions to take place. Although this development has fuelled renewed interest in photochemical



**Figure 1 | Light-controlled stereoselectivity.** Huo *et al.*<sup>1</sup> report a general reaction in which a light-activated iridium catalyst controls the stereochemistry of the product. In this example, an acyl imidazole forms an enolate (blue), which in turn forms a complex with the catalyst (red). Bonds shown in bold project above the plane of the page, whereas hashed bonds project behind the page. The catalyst also converts  $\alpha$ -bromo-2,4-dinitrotoluene (a benzyl halide compound) into a radical intermediate and a bromide ion ( $\text{Br}^-$ ) by donating an electron ( $e^-$ ). The radical reacts only at the top face of the enolate, because part of the catalyst blocks the bottom face. This ensures that the stereochemistry of the product is well defined (the green group in the product projects above the page in most of the formed molecules, rather than below). Me, methyl; Ph, phenyl; *t*-Bu, tert-butyl ( $\text{C}(\text{CH}_3)_3$ , a highly bulky group); Ir, iridium; Br, bromine; the dot on the radical indicates a single electron.



synthesis, control over the three-dimensional structure of the organic products has remained a problem.

This is an important problem, because the ability to form one mirror-image isomer (stereoisomer) of a molecule in preference to the other has profound ramifications in biological and pharmaceutical contexts: the two mirror-image forms often have drastically different physiological effects. Similarly, the macroscopic physical properties of polymeric organic materials can be strongly affected by the stereochemistry of their monomeric components. Stereoselective synthesis therefore remains one of the central challenges in modern synthetic chemistry.

Since 2009, Eric Meggers' research group has been developing methods for preparing ruthenium and iridium complexes as single stereoisomers<sup>3</sup>. In studying these complexes as catalysts for several organic reactions, Meggers and co-workers have demonstrated<sup>4,5</sup> that the three-dimensional arrangement of the complexes can be transferred with exceptional fidelity to the organic products that they create. The same research group — Huo *et al.* — now shows that these transition-metal catalysts are also photoactive, and that this property can be exploited to perform highly stereoselective photochemical reactions.

As a model system, the authors chose to study the  $\alpha$ -alkylation of carbonyl compounds — a benchmark reaction in stereoselective synthesis (Fig. 1). The iridium catalyst first binds to an acyl imidazole compound, creating a structurally well-defined enolate complex. Photoexcitation of this complex initiates an electron-transfer process that converts a reagent (a benzyl halide) into a highly reactive radical intermediate. The geometry of the catalyst shields one face of the planar enolate from reaction (the bottom face in Fig. 1) and forces the radical to form a bond to it preferentially from the opposite face. The iridium catalyst thus serves two distinct roles: it simultaneously photoactivates one component of the reaction (the benzyl bromide) and controls the facial selectivity of the other (the enolate).

These findings will attract considerable attention from synthetic chemists. Photochemical activation typically produces highly reactive intermediates whose stereochemical preferences have historically proved difficult to control<sup>6</sup>. Some of the most successful approaches have needed two catalysts, with one performing the photochemical activation and the other dictating the stereoselectivity of the reaction<sup>7,8</sup>. The discovery of a single transition-metal catalyst that fulfils both roles is a crucial conceptual step forward.

Huo and colleagues' reaction design combines the previously reported, precise stereoselective control exerted by their transition-metal complexes with the practicality of using visible light for photochemistry. Future investigations will surely build on this

impressive result. Because the products of the reported reaction could also be made by more-conventional methods, the next step will be to show that the new catalytic strategy is applicable to other classes of photoreaction for unmet synthetic applications. More broadly, this work provides inspiration for chemists to further explore how photochemistry might be used to transform organic synthesis. ■

**Kazimer L. Skubi and Tehshik P. Yoon**  
are in the Department of Chemistry,  
University of Wisconsin–Madison, Madison,

Wisconsin 53706, USA.  
e-mail: tyoon@chem.wisc.edu

1. Huo, H. *et al.* *Nature* **515**, 100–103 (2014).
2. Schultz, D. M. & Yoon, T. P. *Science* **343**, 1239176 (2014).
3. Gong, L., Wenzel, M. & Meggers, E. *Acc. Chem. Rev.* **46**, 2635–2644 (2013).
4. Chen, L.-A. *et al.* *J. Am. Chem. Soc.* **135**, 10598–10601 (2013).
5. Huo, H., Fu, C., Harms, K. & Meggers, E. *J. Am. Chem. Soc.* **136**, 2990–2993 (2014).
6. Inoue, Y. *Chem. Rev.* **92**, 741–770 (1992).
7. Nicewicz, D. A. & MacMillan, D. W. C. *Science* **322**, 77–80 (2008).
8. Du, J., Skubi, K. L., Schultz, D. M. & Yoon, T. P. *Science* **344**, 392–396 (2014).

## CANCER

# Metastasis risk after anti-macrophage therapy

**Blocking the activity of macrophages may delay the spread of cancer. But new findings show that these immune cells can rapidly rebound to tumours after therapy withdrawal, accelerating lethal metastasis in mice. SEE LETTER P130**

IOANNA KEKLIKOGLOU & MICHELE DE PALMA

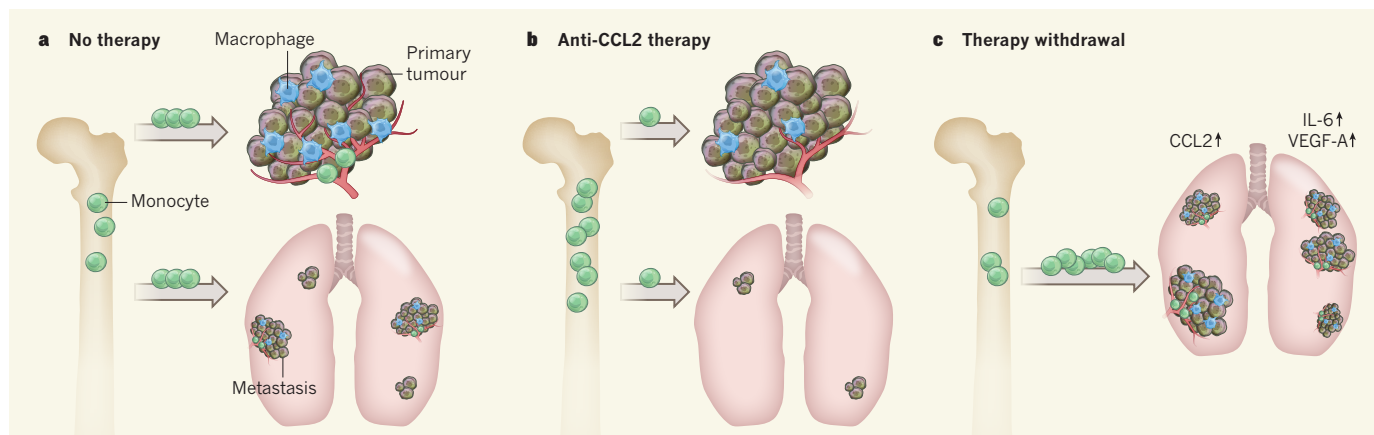
**M**acrophages are immune cells that are key players in our bodies' defence against invading pathogens. Moreover, they participate in organ development, remodelling, healing and disease<sup>1</sup>. Macrophages are also found in tumours, where they seem to support tumour progression and spread by means of several mechanisms<sup>2</sup>. This has prompted the development of drugs that impair macrophage survival<sup>3,4</sup>, block their infiltration into tumours<sup>5</sup> or reduce their pro-tumoural functions<sup>6</sup>. In this issue, Bonapace *et al.*<sup>7</sup> (page 130) report that, in mice, although the continuous blockade of macrophages constrains tumour progression, cessation of the therapy stimulates them to rapidly rebound to the tumours, unexpectedly leading to accelerated metastatic disease.

Monocytes — the circulating precursors of macrophages — enter a tumour from the bloodstream and subsequently differentiate into macrophages<sup>2</sup>. The recruitment of monocytes and their differentiation into tumour-associated macrophages are regulated by signalling molecules released by the tumour or its metastases. One of these is C–C chemokine ligand 2 (CCL2), a protein that attracts monocytes expressing the receptor for CCL2, called CCR2. Blocking the binding of CCL2 to CCR2-expressing monocytes inhibits macrophage infiltration into the metastases that form in the lungs of mice with mammary tumours, delaying the progression of metastatic cancer and extending mouse survival<sup>5</sup>. In humans, both high levels of CCL2 expression

and macrophage infiltration in the tumours correlate with a poor prognosis in some cancer types, such as breast cancer<sup>5,8</sup>. Because tumour metastasis to the body's vital organs is the main cause of cancer-associated death, blocking CCL2 may be an attractive strategy to help combat metastasis in patients with breast cancer and, possibly, other tumour types.

Bonapace *et al.* used a neutralizing antibody against CCL2 to block the protein's activity in mice with mammary tumours. In agreement with previous studies<sup>5</sup>, CCL2 blockade decreased macrophage recruitment to the tumours and reduced the incidence and growth of the lung metastases (Fig. 1). The authors observed that the number of circulating cancer cells shed from the primary tumour, which can travel to the lung and initiate metastases, was reduced during anti-CCL2 treatment. This suggests that CCL2 neutralization had a direct effect on the primary tumour, possibly through macrophage regulation of the growth and characteristics of tumour blood vessels<sup>9</sup> — the first barrier encountered by cancer cells in their journey to distant organs. But the treatment may also affect the establishment and growth of newly settled metastases, for example by inhibiting macrophage production of vascular endothelial growth factor A (VEGF-A), a protein that stimulates the formation of blood vessels in tumours<sup>9</sup>.

But in a dramatic twist, the authors found that interrupting anti-CCL2 therapy accelerated the development of lung metastases and death (Fig. 1). As early as 10 days after withdrawal of the therapeutic antibody, they observed abnormally increased numbers of



**Figure 1 | Monocyte rebounds after therapy.** **a**, The presence of macrophages in primary tumours is associated with tumour growth and metastatic spread to distant organs. (Macrophages differentiate from monocytes, which form in the bone marrow.) **b**, Bonapace *et al.*<sup>7</sup> confirm previous findings that when mice with mammary tumours are treated with anti-CCL2 antibodies, this blocks the egress of monocytes from the bone marrow and impairs macrophage infiltration into primary tumours, leading to reduced formation

of lung metastases. **c**, But the authors also find that cessation of anti-CCL2 therapy is rapidly followed by mobilization of monocytes from the bone marrow to the circulation and their differentiation to macrophages in metastatic tumours, where they promote the formation of tumour blood vessels and tumour growth. This seems to be mediated by macrophage production of the growth factor VEGF-A, and by the signalling molecule interleukin-6 (IL-6).

circulating cancer cells and monocytes in the blood of the mice, and this was associated with rapid infiltration of monocytes in the lung and faster metastatic tumour growth. These pro-metastatic effects were due, at least in part, to the growth-promoting functions of the monocytes recruited to the lung, because removal of the primary tumour after anti-CCL2 therapy also led to increased metastasis, despite the lower numbers of cancer cells in the bloodstream.

So how does cessation of anti-CCL2 therapy promote monocyte rebounds to the metastatic tumours? Bonapace and colleagues suggest that the mechanism may involve heightened CCL2 levels in the lungs of the mice after therapy. Although this phenomenon could be due to the stabilization of CCL2 in complex with the neutralizing antibody, there is clinical evidence that free CCL2 levels surge in patients with cancer who are treated with the human anti-CCL2 antibody carlumab, as early as one week after the first antibody infusion and regardless of subsequent infusions<sup>10</sup>. Together, these findings suggest that pharmacological targeting of CCL2 may trigger a feedback mechanism that fuels CCL2 production. In this scenario, cessation of anti-CCL2 therapy may rapidly (albeit transiently) generate abnormally high levels of free CCL2 that foster monocyte recruitment to the metastatic tumours.

Another mechanism may involve the mode of action of the anti-CCL2 antibody. Systemic neutralization of CCL2 does not impair the production of monocytes in the bone marrow, but rather blocks their mobilization to the circulation. Bonapace *et al.* observed that monocytes accumulated in the bone marrow of their mice during anti-CCL2 therapy, and that therapy withdrawal unleashed these cells, leading to their accumulation in the blood, lungs and tumours. Therefore, post-therapy

rebounds of CCL2 and circulating monocytes may have cooperatively contributed to increasing macrophage infiltration into the metastatic tumours. Once recruited en masse, the macrophages seemed to precipitate metastatic tumour growth mainly through their production of VEGF-A, and, indeed, pharmacological blockade of VEGF-A following cessation of anti-CCL2 therapy restored normal (non-accelerated) tumour progression in the animals.

Anti-macrophage therapies are currently being investigated in patients with cancer, but have not yet received official approval for clinical use. Although carlumab did not show antitumoral activity in initial clinical trials, possibly because it failed to stably neutralize CCL2 in the patients' circulation<sup>10</sup>, other drugs have been developed that specifically target macrophages through different mechanisms. Among these are antibodies or small-molecule inhibitors that block the activity of the colony-stimulating factor-1 receptor (CSF1R), a signalling receptor that controls the differentiation and survival of macrophages<sup>3,4</sup>. Whereas anti-CCL2 antibodies sequester monocytes in the bone marrow and block their recruitment to the tumours, anti-CSF1R antibodies function mainly as monocyte- (and macrophage-) depleting agents<sup>4</sup>, and thus are unlikely to cause monocyte rebounds to tumours post-therapy.

Even so, the depletion of macrophages in tumours may stimulate the intratumoural accumulation of another immune-cell type endowed with protumoral functions, the neutrophil<sup>11</sup>. Neutrophil rebounds in macrophage-depleted tumours may restore, or even precipitate<sup>12</sup>, tumour progression in mice. It remains to be seen whether macrophage-depleting drugs elicit similar compensatory responses in patients with cancer.

Macrophages can suppress the antitumoral functions of T cells of the immune system<sup>2</sup>,

so their transient depletion in tumours may increase the efficacy of immunotherapy designed to evoke T-cell-mediated antitumour responses<sup>6</sup>. Also promising are pharmacological approaches that can 'reprogram' macrophages from being pro- to antitumoral effector cells, by, for example, unleashing their ability to kill cancer cells or to present tumour antigens to T cells<sup>6</sup>.

Bonapace and colleagues' findings underscore both the therapeutic potential and the possible shortcomings of anti-macrophage approaches for cancer therapy. The arsenal of macrophage-targeted drugs is constantly expanding, with increasing sophistication and versatility in their modes of action. When combined with treatments against cancer cells and/or when their use is timed to magnify antitumoral immune responses, this fresh therapeutic asset should prove useful in the enduring fight against cancer. ■

**Ioanna Keklikoglou and Michele De Palma** are at the Swiss Institute for Experimental Cancer Research, School of Life Sciences, École Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland. e-mail: michele.depalma@epfl.ch

- Wynn, T. A., Chawla, A. & Pollard, J. W. *Nature* **496**, 445–455 (2013).
- Noy, R. & Pollard, J. W. *Immunity* **41**, 49–61 (2014).
- DeNardo, D. G. *et al. Cancer Discov.* **1**, 54–67 (2011).
- Ries, C. H. *et al. Cancer Cell* **25**, 846–859 (2014).
- Qian, B.-Z. *et al. Nature* **475**, 222–225 (2011).
- De Palma, M. & Lewis, C. E. *Cancer Cell* **23**, 277–286 (2013).
- Bonapace, L. *et al. Nature* **515**, 130–133 (2014).
- Saji, H. *et al. Cancer* **92**, 1085–1091 (2001).
- Squadrito, M. L. & De Palma, M. *Mol. Aspects Med.* **32**, 123–145 (2011).
- Pienta, K. J. *et al. Invest. New Drugs* **31**, 760–768 (2013).
- Pahler, J. C. *et al. Neoplasia* **10**, 329–339 (2008).
- Swierczak, A. *et al. Cancer Immunol. Res.* **2**, 765–776 (2014).

This article was published online on 22 October 2014.





### Cover image

Andreas Buerkert  
Zakari Soumana

### Editor, *Nature*

Philip Campbell

### Publishing

Richard Hughes

### Production Editor

Jenny Rooke

### Art Editor

Nik Spencer

### Sponsorship

Reya Silao

### Production

Ian Pope

### Marketing

Steven Hurst

### Editorial Assistant

Melissa Rose

The Macmillan Building  
4 Crinan Street  
London N1 9XW, UK  
Tel: +44 (0) 20 7833 4000  
e: nature@nature.com



nature publishing group

We are living in the Anthropocene epoch, the period of time in which human actions have a dominant influence on many of Earth's physical and biological processes. These processes, which are organized into ecosystems, are, in turn, responsible for providing humanity with many essential goods and services. It is therefore important that we rein in our impacts so that ecosystems can operate in a sustainable way, without severe loss or change of function.

This Insight explores the two-way relationship between humanity and natural ecosystems, addresses how it can be managed sustainably and illustrates it with examples from several crucially important systems.

One of the most obvious services that ecosystems provide is food. Graeme Cumming and his co-authors look at terrestrial agriculture and how it has shaped recent societal changes, such as urbanization. The sustainable management of agriculture is a major challenge in a world that is increasingly urbanized and in which most people's lives are far removed from the fields that produce their food.

Even an urbanized world cannot isolate itself from what may at first seem to be detrimental natural processes. Max Moritz and his colleagues discuss the three-way relationship between societies, ecosystems and fire, and how it can be put on a sustainable footing.

Finally, James Watson, Nigel Dudley, Daniel Segan and Marc Hockings look at one of the most successful approaches to curbing human influence and preserving natural ecosystems: protected areas. In both marine and terrestrial systems, protected areas can allow ecosystems to provide goods and services sustainably, and maintain global biodiversity. The authors discuss how well such areas are actually performing and what actions need to be taken to maintain them.

**Patrick Goymer**  
*Senior Editor*

## CONTENTS

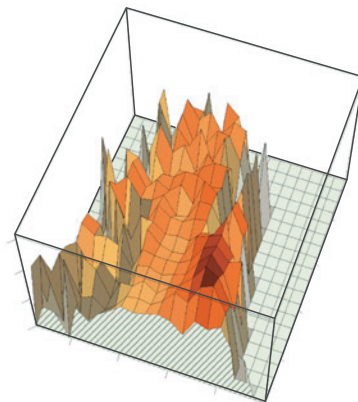
### REVIEWS

#### 50 Implications of agricultural transitions and urbanization for ecosystem services

Graeme S. Cumming, Andreas Buerkert, Ellen M. Hoffmann, Eva Schlecht, Stephan von Cramon-Taubadel & Teja Tschernk

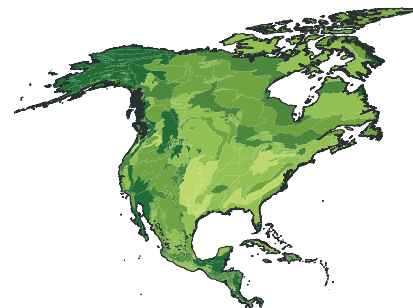
#### 58 Learning to coexist with wildfire

Max A. Moritz, Enric Batllori, Ross A. Bradstock, A. Malcolm Gill, John Handmer, Paul F. Hessburg, Justin Leonard, Sarah McCaffrey, Dennis C. Odion, Tania Schoennagel & Alexandra D. Syphard



#### 67 The performance and potential of protected areas

James E. M. Watson, Nigel Dudley, Daniel B. Segan & Marc Hockings



# Implications of agricultural transitions and urbanization for ecosystem services

Graeme S. Cumming<sup>1</sup>, Andreas Buerkert<sup>2</sup>, Ellen M. Hoffmann<sup>2</sup>, Eva Schlecht<sup>3</sup>, Stephan von Cramon-Taubadel<sup>4</sup> & Teja Tscharntke<sup>5</sup>

**Historically, farmers and hunter-gatherers relied directly on ecosystem services, which they both exploited and enjoyed. Urban populations still rely on ecosystems, but prioritize non-ecosystem services (socioeconomic). Population growth and densification increase the scale and change the nature of both ecosystem- and non-ecosystem-service supply and demand, weakening direct feedbacks between ecosystems and societies and potentially pushing social-ecological systems into traps that can lead to collapse. The interacting and mutually reinforcing processes of technological change, population growth and urbanization contribute to over-exploitation of ecosystems through complex feedbacks that have important implications for sustainable resource use.**

Contemporary research suggests that humanity is over-exploiting the environment<sup>1</sup>, driving global climate change, eutrophication, degradation of ecosystems and biodiversity loss<sup>2</sup>. At the same time, the world's human population is projected to grow from 7.2 billion people to 9.6 billion by 2050 (ref. 3). Although most agro-ecosystems have coped with anthropogenic pressures<sup>4</sup>, we cannot assume they will continue to meet our increasing demands<sup>5</sup>. Food-production systems are now global, with attendant benefits and risks<sup>6</sup>; the diversity of farmed crops is declining<sup>7</sup>; and environmental degradation from agriculture is widespread<sup>8–10</sup>. These trends are eroding the resilience of agro-ecosystems to anthropogenic perturbations such as climate change<sup>6,11,12</sup>.

Reconciling the demands of the growing human population with ecological sustainability is increasingly difficult<sup>13</sup>. The Millennium Ecosystem Assessment<sup>14</sup> classified ecosystem goods and services (ESS) into four categories: provisioning, regulating, supporting and cultural. It also acknowledged that ecosystems can provide or contribute to disservices, such as pathogens and floods. Subsequent analyses have generally focused on single services, or on ESS as outcomes of ecosystem-focused or food-production-focused models<sup>15–17</sup>. The underlying drivers of ecosystem degradation are, however, economic activities that are not themselves ecosystem-focused<sup>18</sup> and may be separated from their own consequences by long socioeconomic supply chains<sup>19</sup>. ESS research has concentrated on ecosystems<sup>20</sup>, rather than the institutional, political and socioeconomic drivers of ecological change<sup>21</sup>. Even the recognition that monitoring ESS requires not only ecological but also socioeconomic data is relatively recent<sup>22</sup> and has not yet influenced the ways in which important policies, such as international trade agreements and development goals, are designed and implemented<sup>22,23</sup>.

Social-ecological systems are complex and adaptive, and attempts to manage them often have unintended consequences<sup>11,24</sup>. To manage ESS sustainably, we need to understand the trajectories of change that have produced our current situation and continue to shape it; the interactions, feedbacks and trade-offs between different services and the social-ecological interactions that produce them; and the ways in which fundamental structural changes (those that require new system models, rather than simply adjustments to existing models) occur within the ESS context. Developing this understanding requires us to connect people and ecosystems in an interdisciplinary social-ecological systems

framework<sup>25,26</sup>. We address this challenge by proposing a simple conceptual model that shows how a systems perspective on ESS, in the context of agricultural transitions and increasing urbanization, helps to explain ecological over-exploitation.

## Service shifts in agricultural transitions

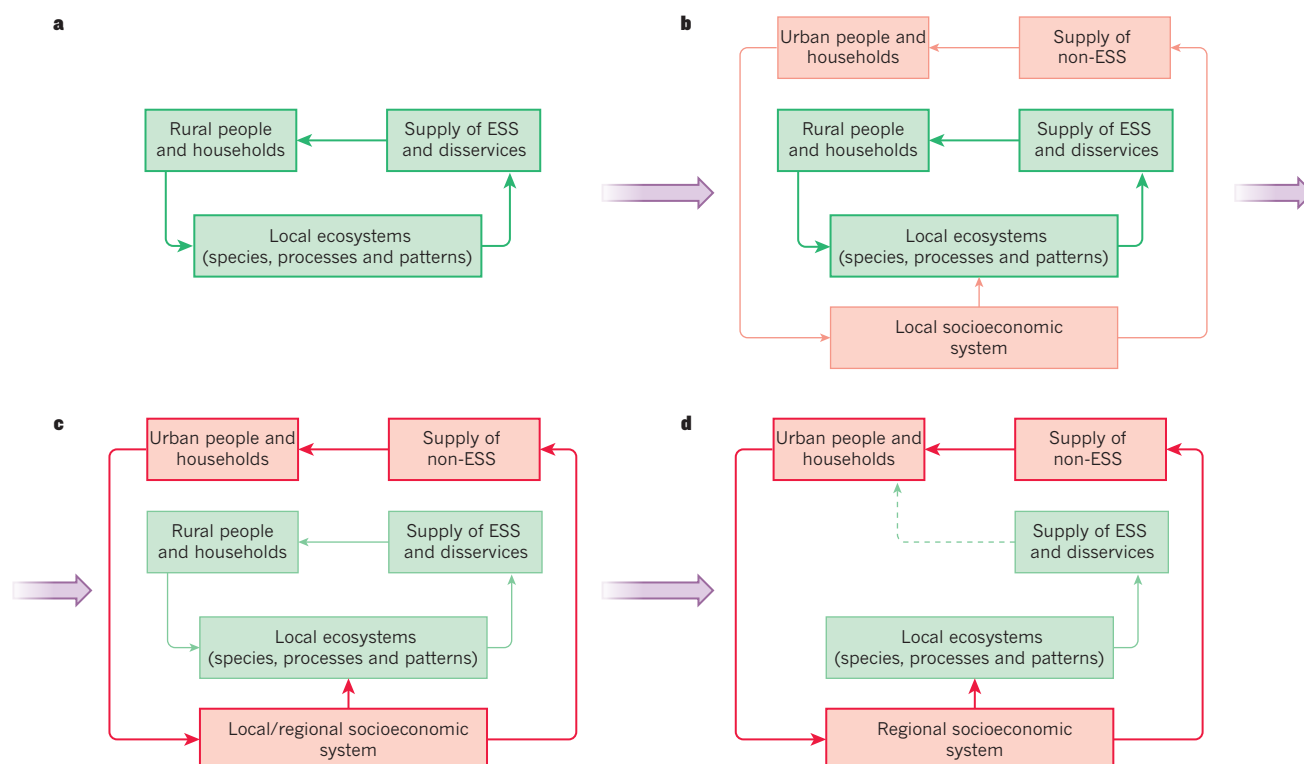
In a sense, human history is the story of the great transition from hunting and gathering, through farming, to the present situation in which less than 1% of the population (in many industrialized countries) is directly employed in food production<sup>27–29</sup>. In Germany, for instance, the average farmer in 1950 fed 10 individuals, but in 2010 he or she was feeding 131 people<sup>30</sup>. Despite its importance, there have been few attempts to develop formal social-ecological models of this transition. Generic systems models of ESS are also surprisingly hard to find. Most empirical analyses apply to individual goods or services, such as water flows or climate regulation<sup>21</sup>. A few published studies have quantified trade-offs between different ecosystem services<sup>31–35</sup>, but we are unaware of any formal, causal systems models that provide a broad overview of ESS across multiple categories and scales. Despite the existence of a wide range of land-cover change models<sup>36</sup>, and a growing interest in transforming cities for greater environmental sustainability<sup>37</sup>, the changes in service provision that are likely to happen during the transition from an agricultural society to an industrial society are poorly specified.

If a human population is stable and most of the people depend directly on ESS, the feedbacks from ESS to human well-being are clear. Many cultures have developed rules and traditions that, under normal conditions, maintain their own resource base (for example, Balinese water temples as irrigation systems<sup>38</sup>; the protection of sacred forests in southwest Madagascar<sup>39</sup>; and the release of trained eagles, once they reach 5 years old, back to the wild by the Kazakh Golden Eagle hunters of western Mongolia<sup>40</sup>). Although rules for natural resource management are not always effective, if a local equilibrium between resource use and human population size is maintained, a 'green loop' that avoids long-term degradation of ecosystems can be sustained.

The green loop starts to break down when human populations grow as a result of technological change that increases food supply and life expectancy. Population density and infrastructure increase as urban settlements create alternative livelihood opportunities, provide security

<sup>1</sup>Percy FitzPatrick Institute, DST/NRF Centre of Excellence, University of Cape Town, Rondebosch, Cape Town 7701, South Africa. <sup>2</sup>Organic Plant Production and Agroecosystems Research in the Tropics and Subtropics, Universität Kassel, Steinstr. 19, D-37213 Witzenhausen, Germany. <sup>3</sup>Animal Husbandry in the Tropics and Subtropics, Universität Kassel and Georg-August-Universität Göttingen, Steinstr. 19, D-37213 Witzenhausen, Germany. <sup>4</sup>Department of Agricultural Economics and Rural Development, Georg-August-Universität, Platz der Göttinger Sieben 5, D-37073 Germany. <sup>5</sup>Agroecology, Georg-August-Universität Göttingen, Grisebachstr. 6, D-37077 Göttingen, Germany.





**Figure 1 | The green-loop to red-loop transition.** In this transition, as the population grows, the red loop overwhelms the green loop to become the dominant regime driving the use of ecosystem goods and services (ESS). **a**, In the starting green loop, rural populations manage their local ecosystems. **b**, As the population grows, a 'shadow' red loop begins to develop; changes in socioeconomic variables, such as increased demand for food, fibre and fuel, lead to greater local ecosystem impacts. **c**, The red loop gains prominence as

demand for services shifts from a need for ecosystem services to a need for non-ecosystem services. **d**, As the demand for services shifts, the red loop becomes the dominant driver in the flow of ecosystem services and is accompanied by an upscaling and specialization process that results in the gradual alienation of urban people from the ecosystem; the strength of the connection between the local ecosystem and society is heavily reduced (dashed line). This can easily result in over-exploitation of ecosystems and ecological degradation.

and increase economic, social and political complexity. However, urban dwellers typically have less contact with their primary resource base<sup>41</sup>. Over time, the ability of local ecosystems to supply a full range of ESS to growing settlements is reduced by one or more possible causes. First, the area required to meet the needs of each family exceeds the boundaries of the area that they cultivate, and the needs of the entire population exceed the resources that they can access directly. The local ecosystem cannot produce enough food, particularly during periods of adversity. In addition, as settlements grow, the surrounding ecosystems are increasingly modified for provisioning services such as food and water, often at the expense of other kinds of ESS. Second, increasing population density makes simple forms of waste disposal impractical, necessitating technological solutions. Third, as the settlement grows the logistics of access to ESS and travel time make it impractical for each household to extract everything that it needs from the local ecosystem. This creates demands for trade, technology and infrastructure to enhance resource-use efficiency (particularly of land and water), and the need for specialized production roles (for example, farmer or blacksmith)<sup>42</sup>. Last, an institutional environment that enables specialization and exchange, as well as the planning and maintenance of public infrastructure, requires individuals such as administrators, merchants and law-enforcers who do not contribute directly to food production, further distancing individuals from ecosystems. As societies find solutions to these challenges, local economies and populations grow and the tasks that people perform become more specialized<sup>42</sup>. These changes gradually transform a system in a green loop to one in a 'red loop' (Fig. 1) as three trends unfold. First, demands for non-ESS continue to increase, resulting in changes in institutions (rules, laws, policies and customs) and governance systems as well as the construction of infrastructure (housing, roads and reservoirs). Second, urban settlement

and specialization foster technological progress, which is then pursued systematically. Technological progress in agriculture, specifically, can strengthen both population growth and urbanization<sup>43,44</sup>. Third, because provisioning from local ecosystems can no longer meet local demand, many needs that were formerly met by local ecosystems are outsourced, resulting in an increase in the geographic extent of supply and demand (upscale). This trend is reinforced by developments in transportation technology and the demand for foreign products (for example, spices and precious metals) by a growing and increasingly wealthy population. Thus, a gradual transition occurs from an economy based on ESS to one based on non-ESS and remote extraction. In the process, the perceived importance of ecosystems to people decreases. The proportion of people who extract goods directly from ecosystems (farmers, fishers and loggers) declines and their status might be reduced. During the transition period, which may last for decades, elements of both red and green loops coexist (Fig. 2). Typically, this increases socioeconomic diversity, spatial heterogeneity and inequity, often with the formation of spatial gradients in ecosystem service provision and socioeconomic variables related to proximity to various resources<sup>45</sup>.

The socioeconomic dynamics that are driven by growing markets for non-ESS, together with upscaling, ongoing technological change and related acceleration of population growth, have many consequences for ESS. Society's ecological footprint grows<sup>46</sup>. As people's reliance on ecosystems becomes less obvious, they become less aware of ecological degradation and less concerned about it. They might also be too overwhelmed by local change to pay attention to their regional and global impacts<sup>47</sup>. As the connections between food production and food consumption (as well as feed and fuel production and use) become less apparent, societies unintentionally place increasing pressure on dwindling resources. In addition, the ability of people to censure others for

abusing natural resources declines<sup>48</sup> because social interactions between producers and consumers weaken.

Once the transition from green-loop to red-loop dynamics is underway, the red loop becomes the dominant driver of societal change. Institutions and actions that conserve ESS and contribute to their sustainability must then be negotiated in new action forums<sup>49</sup> in which many powerful and often competing actors, such as politicians, mining corporations and manufacturing industries, push to enhance the provision of non-ESS, often at the expense of ESS. The shift from green-loop to red-loop dynamics thus underpins a gradual regime shift<sup>24,50</sup> in the entire social–ecological system.

Although the transition occurs gradually, the shift from a green to a red loop represents a fundamental change in system functioning that requires two different system models, rather than parameter changes within a single model. The two loops are alternate social–ecological states, each of which has reinforcing feedbacks that buffer it from change. The key slow-changing variables in the system<sup>51</sup> are increasing human population and population density, which create amplifying feedbacks that rapidly ratchet up the demand for ESS and non-ESS; technological change, which accelerates population growth and enables a growing proportion of people to obtain their livelihoods in ways unrelated to agriculture; and a loss of biodiversity, which can lead to eventual socioeconomic collapse.

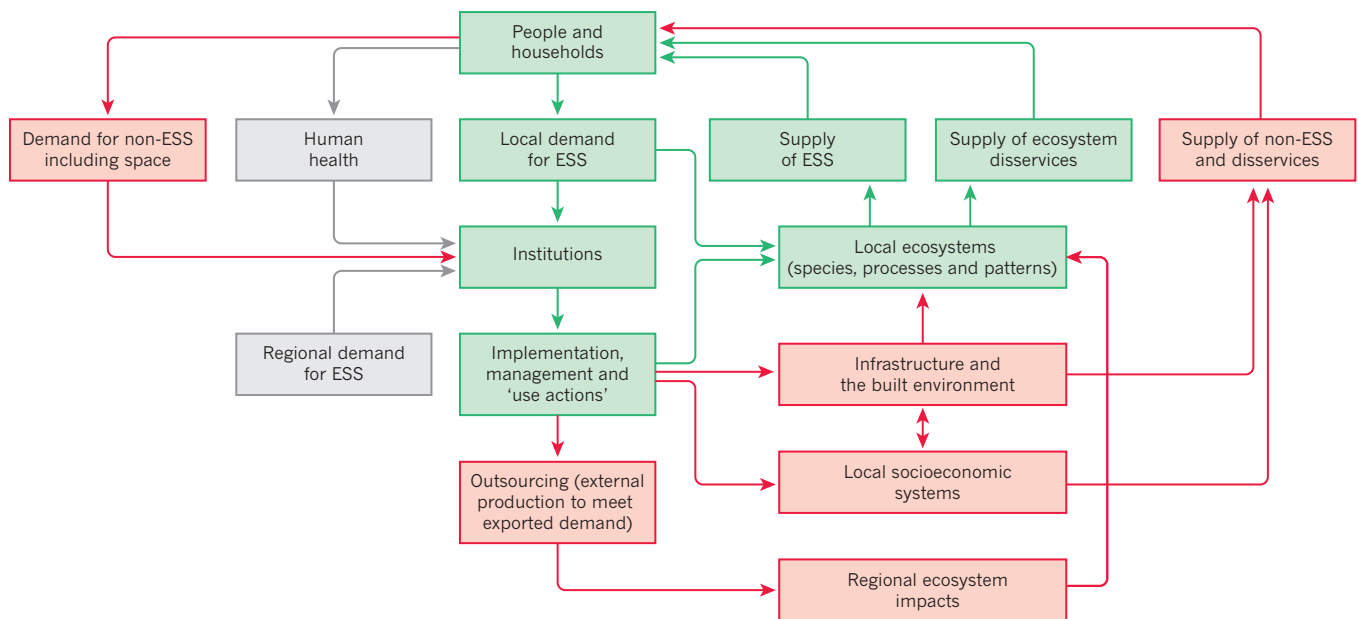
During transition, the proportion of aggregate income obtained directly from ecosystems declines from high (green loop) to low (red loop)<sup>52,53</sup>. A shift occurs from high to low relative prices for basic ESS (provision of staples such as wheat, potatoes and cassava) and to higher demand for special commodities (luxury food items such as a range of fruits)<sup>54</sup>. The value of sustaining and regulating services also increases while the value of provisioning services declines. For example, access to electricity and fossil fuels in cities reduces the reliance on local wood-fuel production<sup>55</sup>. Demands for cultural services might change with peoples' perception of nature, and willingness to pay for cultural services may increase as natural landscapes become

scarcer. For example, temples and shrines in heavily populated cities in Japan and Thailand have found new significance as places in which to experience nature<sup>56</sup>. Upscaling and increased trade in a red-loop population are not necessarily unsustainable if they lead to an equilibrium between the human population and resources at larger geographic extents.

Lock-in to a red loop need not be an 'end point'<sup>57</sup>. Demand for new cultural services, such as walking trails and ecotourism, could lead to a reintegration of local ESS into urban economies and politics (although ecosystems may be species-poor by the time this occurs). Concerns for human health could also lead to measures to prevent environmental degradation. Ageing and declining post-peak human populations will bring new dynamics and possibly, if sufficient biodiversity remains, the potential to return to more direct interactions with ESS. It remains unclear, however, whether efforts to 're-green' cities (for example, through urban rooftop gardens<sup>58</sup>) can persist as the human population continues to grow, and whether cities will become unsustainable without efforts to make them greener and more self-sufficient.

The transition from green- to red-loop dynamics occurs through feedbacks between technological change, population growth and ecosystem change. The resulting red loop has the potential to sustainably reconcile these forces by solving service supply and distribution problems. There may, however, be hurdles that prevent a successful transition and/or reduce the sustainability of the red loop. Overconsumption in the red loop and failure to regulate ecological decline can produce a 'red trap'. Rural poverty and ecological degradation in the green loop may reinforce each other, leading to a 'green trap'. In both cases, systems must reorganize or they will collapse (Fig. 3).

Tests of our model require long-term time series data for agricultural production, demography, economic developments and ecosystem change (Table 1). As a first step towards grounding the model empirically, we review evidence from three case studies: Sweden, as an example of a green-loop to red-loop transition; the Sahel, focusing



**Figure 2 | Detailed interactions and feedbacks during the transitional period between green and red loops.** Basic household needs create a local demand for ecosystem goods and services (ESS). This may be expressed as direct and unregulated impacts on ecosystems, or, more typically, as 'use actions' (consumptive and non-consumptive) that are governed by rules, laws, policies and customs (institutions). Among use actions, those that have the highest ecological impacts are generally those that involve direct extraction of resources (for example, logging, cultivation or water extraction). Use actions affect the provision of ESS as well as 'disservices' (pathogens, crop damage

or floods). The degree to which human needs are met by ecosystem services then affects future demand, completing the loop. The direct interactions of people and ecosystems are gradually overrun by the red loop, in which the focus is non-ESS, despite the continued importance of ecosystems for the community. Ongoing local and regional impacts on ecosystems are hidden from urban dwellers by outsourcing and infrastructure development. The two 'wild card' variables (in grey), human health and regional processes, may be present in either red- or green-loop situations and may create ecological or socioeconomic surprises that can alter system dynamics.



on Niger, as an example of a green-loop to green-trap transition; and Beijing, as an example of a red-loop to red-trap transition (in the absence of an unequivocal contemporary example of a country in a red trap).

### Green loop to red loop in Sweden

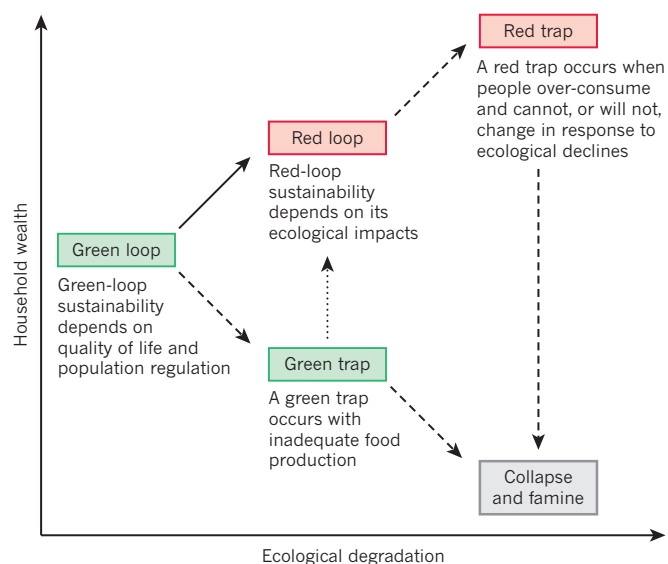
For more than 1,000 years, Sweden had low population levels and a dominantly agrarian lifestyle, consistent with a green-loop dynamic. It still has one of Europe's lowest population densities (around 9.5 million inhabitants; 21 people km<sup>-2</sup>), but between 1750 and 1850 the population doubled and its subsequent growth was much faster<sup>59,60</sup>. Around 1870 to 1890, population growth triggered a switch from a green loop (or possibly even a green trap; more than 1% of the population emigrated to America every year during the 1880s<sup>61</sup>) to a red loop. Rapid economic development, fuelled by engineering, mining, and the steel and pulp industries as well as internal institutional changes and a growing export market, took place between 1870 and 1914 (ref. 62). Two world wars and a global recession reduced economic growth, but gross domestic product (GDP) grew rapidly after 1950 (ref. 60).

Since 1950, industries and business services have expanded, whereas agricultural production has remained relatively constant. Infrastructural assets — buildings and machinery — grew by two orders of magnitude from 69 billion year-2000 Swedish kronor in 1965 to 2.5 trillion year-2000 Swedish kronor in 2000. By contrast, employment in agriculture declined from nearly a million people in 1880 to under 50,000 in 2000, and roughly 20% of agricultural land was removed from production between 1949 and 1999 (ref. 63). Sweden managed to retain substantial natural resources through its agricultural transition. In 2013, 69% of the country was forested, 8% was used as agricultural land and only 2.8% was 'built-up'<sup>64</sup>. Transport infrastructure (roads, railways, harbours and airports) accounted for 40% of built-up land, with total infrastructure occupying a greater area than residential dwellings.

Advances in technologies and farming methods seem to be helping Sweden to remain sustainably within a red loop, with reductions in local environmental degradation and stable or increasing food production. According to Statistics Sweden<sup>64</sup>, total household water withdrawals between 1995 and 2010 declined from 616,000 m<sup>3</sup> to 576,000 m<sup>3</sup> per year; and for agriculture, from 137,000 m<sup>3</sup> to 99,000 m<sup>3</sup>. Nitrogen inputs into water bodies declined from 34,527 t in 1995 to 24,416 t in 2005 and for phosphorus, from 970 t to 733 t during the same period. From 1965 to 2012, the area farmed and the numbers of individual people engaged in farming strongly declined, but yields per hectare of wheat (summer and winter) increased from 6,880 kg ha<sup>-1</sup> to 11,110 kg ha<sup>-1</sup> and annual production more than doubled from 1,039,320 t to 2,289,300 t.

In a red loop we expect a disconnect between people and local ecosystems, with negative consequences for biodiversity and ESS. The available evidence supports this view. The greatest losses of Swedish grasslands, one of the country's most species-rich habitats, occurred before 1950 and created an extinction debt for habitat-specialized vascular plants, with species still being lost from the remaining grasslands<sup>65</sup>. Extensive loss of old-growth forest dates back to a 1948 policy that enabled clear-cutting and over-use of herbicides in Swedish forests. The remaining forests of high conservation value are considered too small and too fragmented to meet current forest and environmental policy goals<sup>66</sup>. The Swedish Environmental Protection Agency was only created in 1967, and biodiversity conservation has only been a nationally agreed objective of forest management since 1992 (ref. 66). In the Baltic Sea, the Swedish cod harvest peaked at 59,000 t in 1984, but had dropped to 16,000 t by 1993 as the fishery collapsed<sup>67</sup>. A review of the impacts of agricultural intensification on essential ESS in Sweden between 1950 and 1999 (ref. 63) found that most of the measures indicated a loss of ESS from the Swedish agricultural landscape. These included a 60% decline in native pollinator abundance and 46%, 33% and 14% increases in the concentrations of the heavy metals mercury, cadmium and lead, respectively.

Sweden has met many of its needs by upscaling, as predicted by our



**Figure 3 | States, traps and transitions along the rural to urban gradient.** A typical development trajectory from an agricultural (green loop) to industrial (red loop) society involves individual households gaining wealth while some level of ecosystem degradation occurs. Depending on population growth rates and governance, societies may grow without true socioeconomic restructuring (green trap) or become rich and continue to over-exploit ecosystems (red trap). The dashed lines indicate avoidable transitions. Both traps can lead to socioeconomic collapse. One of the primary challenges of development and policy initiatives is to shift societies from a green trap to a red loop (dotted line) without heavily altering consumption patterns, thus maintaining a relatively high individual quality of life without entering a red trap.

model. The ecological footprints of large cities in the Baltic Sea region for food and timber production and waste assimilation are more than 565–1,130 times their combined area<sup>68</sup>. Swedish imports and exports increased sevenfold from 1975 to 2000, with unknown ecological impacts on remote locations. Internal biodiversity loss or an external limit to growth (such as climate change) may yet affect Sweden's economy<sup>69</sup>. We could, however, find no obvious evidence that Sweden has entered a red trap. For the moment, it is an example of a shift from a green to a red loop that first increased and then reduced the impacts of the growing human population on local ecosystems.

### Green loop to green trap in the Sahel

In Niger, millennia-old, environmentally specialized societies of pastoralists, agro-pastoralists, fishermen and traders indicate the long-term adaptations of people to ecosystem limitations and opportunities. In some parts of the pre-colonial Sahel, the wealth created as a result of labour division and the inter-regional trade of gold, salt and slaves led to the formation of cultural centres such as Timbuktu and Djenné in Mali between the thirteenth and sixteenth centuries<sup>70</sup>, proving the economic success of combined trade, regional migration and agro-pastoralism in successfully defying unpredictable rainfall. The existence of a relatively sparse rural society in the Sahel for several thousand years suggests a stable green loop.

The slave trade during the eighteenth and nineteenth centuries resulted in the loss of up to 3 million African inhabitants, affecting the workforce and cultural progress. Population recovery during the twentieth century rapidly led to a shortage of fertile land. Together with erratic rainfall, low soil fertility has, for centuries, limited the effectiveness of agricultural intensification efforts<sup>71</sup>. Shorter fallow periods have led to the expansion of cropping systems into ever more marginal drylands. The resulting large drop in per capita cereal production has required rapidly increasing cereal imports<sup>72</sup> (Fig. 4). Although between 1970 and 2012 the area of harvested cereals expanded from 2.3 million to 10 million ha, cereal imports increased from almost zero to 340,000 t (Fig. 4).

**Table 1 | The main premises (both well proven and those for which the evidence is circumstantial) underpinning our model, and forms of evidence on which proof or disproof of our argument rests**

Model stage, prediction or hypothesis	Evidence that would support the model	Relevant data
Relatively stable populations of low densities are maintained with an agrarian or pastoral lifestyle.	Lower population density before the formation of cities.	People per hectare before urbanization, showing evidence of stability in numbers.
Low population density is, or was, ecologically sustainable over timescales of centuries.	Low population density did not lead to degradation of ecosystem services (in addition to evidence of more than 250,000 years of human existence in Africa).	Estimates of how much land was needed for sustainability, for example, the number of hectares per household needed to maintain shifting agricultural system productively for more than 50 years, and proof that this much land was available.
Population increase leads to an increase in the number and size of cities (and/or land degradation and poverty).	Increasing urbanization, declining per capita agricultural production, declining household smallholding sizes as well as intensification as a temporary fix, or failure of agricultural production to sufficiently increase to meet demand.	City sizes, urban population demographics and urban growth rates; per capita production of key food crops; and village, farm or smallholding sizes.
In cities, the proportion of household income from agriculture drops as the society enters a red loop. The agricultural transition divides urban (red loop) and rural (green loop) people.	Differences in household income sources between rural and urban dwellers, declines in proportion of income from agriculture (as an income source) or the increasing role of non-ESS.	At the microscale, household-level data on net income and sources of income; at the macroscale, agricultural production as a proportion of GDP between urbanized and developing countries; and data on service industries and government or city expenditures.
Once in a red loop, upscaling of production systems must occur to meet the food demands of the urban population.	Upscaling, for example, greater ecological impacts on the surrounding countryside, impacts of urban demand on rural production systems and markets, and increased importance of trade.	Data on food prices, diversity and demand from city dwellers (compared with rural dwellers); rates of land conversion around cities; and per hectare production of crops in relation to market growth.
Red-loop dynamics reduce the connections of city dwellers to the countryside, fostering further ecological degradation.	Increasing rates or magnitudes of ecological impacts as urbanization levels increase, with less obvious dependency by city dwellers on provisioning ecosystem services, and increased ignorance about ecosystems (for example, where food comes from or what natural habitats really look like).	Data on land-cover change and biodiversity loss as urbanization occurs, ideally compared with a dysfunctional green-loop situation (increasing population and declining quality of life).
Existence of a green trap.	Population increase is possible without urbanization (or the total population may grow more rapidly than the urban population).	Data showing increasing rural population and declining per capita production.
Existence of a red trap.	Unsustainable consumption by wealthy societies.	Despite arguments for the existence of red traps based on archaeological data, because of global upscaling few, or no, clear-cut contemporary examples exist.
Potential for collapse.	The demonstration that collapse is possible from both green-trap and red-trap situations.	Archaeological evidence for social-ecological collapse in past civilizations, both agrarian and urbanized. Contemporary examples are harder to find because of technology and globalization.
Potential for shifting from a green trap to a red loop.	Urbanization and migration can provide a short- or intermediate-term solution to rural poverty.	Data on household incomes for societies (for example, Gini coefficients) as they go through a transition.

GDP, gross domestic product; ESS, ecosystem goods and services

These trends correspond with a shift from a green loop to a green trap, in which poor rural populations remain enmeshed in rural poverty. Apart from during the two big Sahel droughts (in the early 1970s and mid-1980s), Niger has coped with the per capita decline of its rain-fed cereal production by upscaling. In our model, this indicates a shift towards a red loop. However, the economic basis for imports was the uranium boom — recently complemented by revenues from oil and gold — which resulted in an availability of funds without the creation of a full set of economic, infrastructural and institutional assets that would characterize a red loop. The decline in demand for nuclear fuel during the 1990s therefore resulted in a food crisis and political instability.

In response to the green trap, the rural population of Niger migrated. In 1951, the urban population was 6% of the country's 3.3 million inhabitants; by 2012 it was 17% of 16.6 million<sup>72</sup>. As people in Niger attempt to escape the green trap, the intensive production of vegetables in urban and peri-urban agricultural systems and in irrigated gardening systems of southeastern Niger has increased; for example, onion sales in the Maradi region increased from 26,000 t in 1961 to 370,000 t in 2011 (ref. 72). Imports of staple foods, largely financed by foreign aid, have allowed the urban population and its alienation from ESS to continue to grow (upscaling based on external economic resources). Upscaling of demand without expansion of local supply has led to further neglect of the rural sector, putting additional strain on ecosystems, leading to more ecosystem degradation, and making it increasingly difficult to escape the green trap. For example, reliance on wood fuel from

the marginal shrublands that formerly surrounded the cities<sup>73</sup> has led to the widespread loss of vegetation cover and a decline in associated regulating and supporting services.

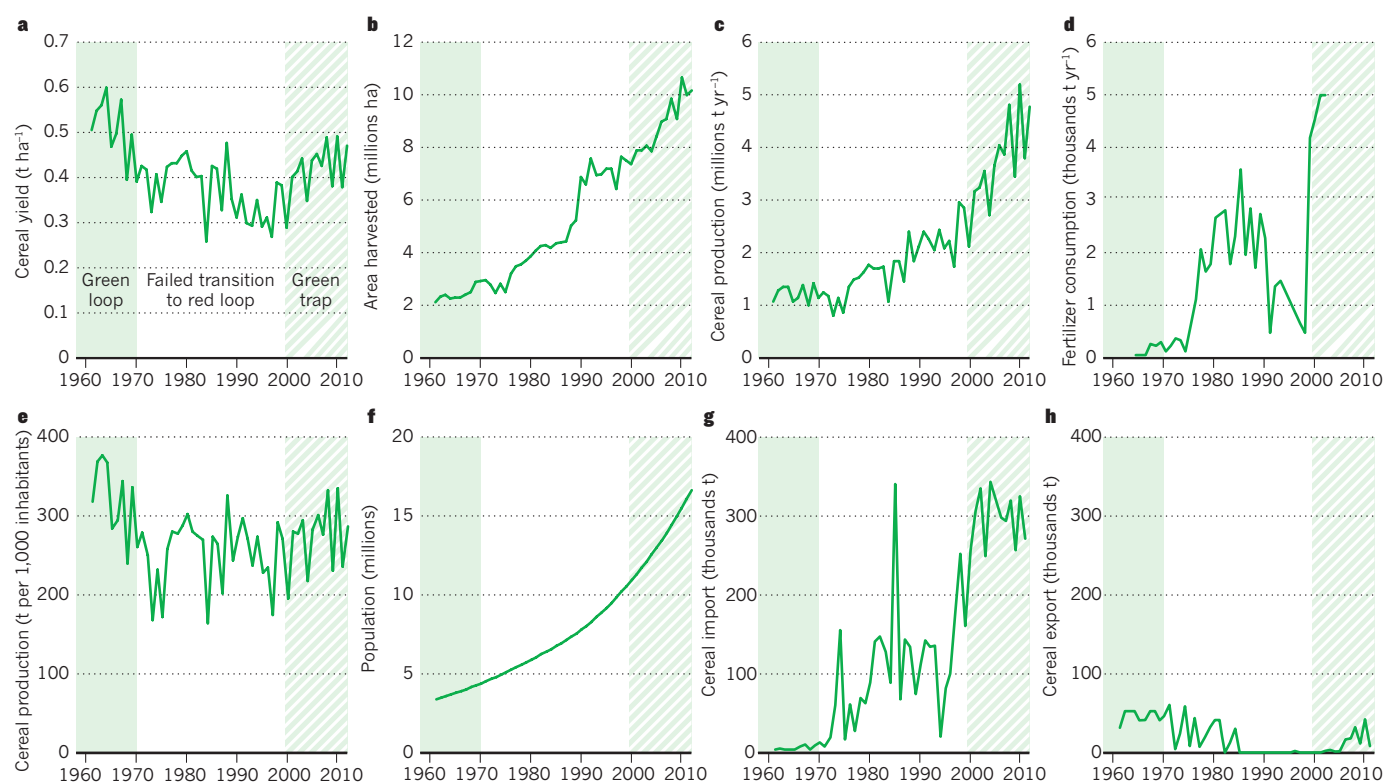
Agricultural innovations proposed for the Sahel over the past 50 years have largely failed because food production is hampered by a combination of climatic unpredictability and political neglect<sup>74</sup>. A few examples from Sudano-Sahelian West Africa<sup>75,76</sup> indicate that agricultural intensification with positive feedback loops to ESS is possible, in principle, in this region. It depends, however, on effective local policies, risk-reducing technologies and stable market demand for commodities that support farmers' investments in agriculture (as well as curbing the present 3.9% per annum population growth rate).

### Red loop to red trap in Beijing

Beijing is situated on the fertile North China Plain. It has an average annual precipitation of 578 mm and relies, for staple foods, on an intensive double cropping system of maize (corn) and wheat. Although the lack of water has limited the development of the Beijing basin area for centuries, and despite China's one-child policy, the greater metropolitan area has grown from 9 million inhabitants in 1978 to more than 17 million in 2009 (ref. 77).

Population growth, exacerbated by immigration, rapid industrialization and changes in consumer demands, has led to an ever-increasing demand for water resources. Beijing's per-capita water-storage capacity of 300 m<sup>3</sup> is 12.5% of China's urban average and 3% of the world's.





**Figure 4 | Development of cereal production in Niger between 1960 and 2014.** Cereal data illustrate the failed transition from a green to a red loop, and the resulting entry into a green trap. Cereal yield per hectare (a) has declined and the large expansion of the cultivated area (b) is primarily responsible for the increase in total cereal production (c), despite an overall

increase in fertilizer use (d). Cereal production per 1,000 people (e) dropped slightly as the human population grew (f). This would have led to a decline in cereal availability per capita. However, since the 1990s, extensive cereal imports (g) have compensated for the shortfall, with the linked collapse of cereal exports (h).

Overuse of ESS is evident: 60% of Beijing's total water use, and 80% of its irrigation water, is fossil groundwater, which is unrenovable. Average water table levels in 2000 were 8.1 m lower than in 1980 and 12.2 m lower than in 1960 (ref. 78). In 2005, agriculture consumed 38% of the total water, for industry the value was 20%, and for municipal and residential purposes it was 39%; the latter is rapidly increasing, leading to fierce competition between these sectors<sup>78</sup>. Around 70% of the irrigation water in the North China Plain is wasted by evaporation, deep percolation or run-off<sup>79</sup>. China's central government has now implemented measures, such as the use of plastic mulching and tree-crop interplanting on large areas, to enhance water conservation.

High-intensity agricultural production in the Beijing area satisfies only 17% of city dwellers' demand for grain and 31% of their demand for vegetables<sup>77</sup>. Heavy environmental contamination has occurred from uncontrolled wastewater discharge into water bodies, nitrate leaching from over-fertilization, and the release of gaseous pollutants and aerosols into the atmosphere. An estimated 75% of urban residents in China live in areas in which the air quality is below the country's own standards; fine particle, emissions of sulphur dioxide and nitrogen oxides and subsequent fallout of acid rain<sup>80</sup> affect an even higher proportion of Beijing's population. In 1997 the nationwide death toll from air pollution was already estimated to be 300,000 people per year<sup>81</sup>. Annual total aerial nitrogen deposition rates in China rose from 13 kg N ha<sup>-1</sup> in the 1980s to 21 kg N ha<sup>-1</sup> in the 2000s, of which agricultural nitrogen sources contributed two-thirds<sup>82</sup>. Recent data for the Beijing area show annual total dry and wet nitrogen depositions of more than 90 kg ha<sup>-1</sup> per year<sup>83</sup>, resulting in widespread acidification of the generally well-buffered surface soils of China's croplands<sup>84</sup>.

In recent decades, per capita income in Beijing has risen faster than the cost of living and the proportion of household income spent on food has declined. Hence, Beijing has witnessed a transition to urban lifestyles, a growing dependence of food markets on distant ESS (upscaling)

and the breakage of direct feedbacks from local ecosystems to the local population. So far, upscaling seems to have been a successful strategy for dealing with a potential red-trap situation. However, it is unclear whether further development in Beijing will be sustainable given ongoing declines in ESS and human well-being<sup>85</sup>.

### General implications

Although ecosystems are the foundation on which non-ESS rest, demands by urban societies for non-ESS make the connections between humanity and ecosystems less obvious and less immediate. The social-ecological dynamics of ESS are strongly driven by the more general demands of society for non-ESS and by the changes in the scales of supply and demand, for both ESS and non-ESS, that accompany the transition from agricultural to industrial societies. The first point in particular has not been incorporated into the ESS literature.

Agricultural transitions are fundamentally linked to human population growth<sup>86</sup>. Growing societies that attempt to remain in a green loop will almost inevitably enter a green trap, which could result in greater biodiversity losses than a red loop<sup>87</sup>. Few contemporary societies exist in a green loop, and those that come closest to doing so are often socially and economically marginalized and vulnerable to external exploitation of their ecosystems<sup>88,89</sup>. Contemporary societies that seem to have best navigated a balance between ecological sustainability and human well-being are those, such as Sweden, that have entered a red loop without shifting exploitation to red-trap levels. The red loop has bought such societies additional time, and the best-case scenario is that socioeconomic feedbacks within the red loop (for example, declining fertility, or simply longer inter-generational times and smaller families) could reduce population growth and ecological footprints before these systems enter a red trap and collapse<sup>90</sup>.

Scale is of critical importance here: the cumulative effect of many local or regional red loops may be a global trap, for example if their

combined greenhouse-gas emissions trigger climate change. We would expect to find scale dependencies in the relative importance of different links in the model. At local extents, questions of access and infrastructure development may dominate red-loop ecological impacts. At national extents, the model may capture the basis of an economy as rural or urbanized, and upscaling can be perceived as globalization. For empirical analyses, we suggest an initial unit of analysis as the household, with aggregation of household data across a range of different spatial and temporal scales, and institutional levels.

Our model shares some elements with the environmental Kuznets curve<sup>91,92</sup>, which suggests that indicators of environmental degradation follow an inverted U-shaped curve over the course of economic development. We do not wish to reinstate Kuznets' hypothesis, which has been criticized<sup>92,93</sup>; but the different pathways that we have identified explain why the environmental Kuznets curve might apply to some societies (such as those undergoing a green- to red-loop transition) and not to others (such as those that are caught in, or heading towards, red or green traps).

It remains unclear whether, how and when ecological debts incurred during industrialization will have to be repaid. Human survival depends on maintaining functional, resilient ecosystems and resulting ESS through the bottleneck of maximum human population. The loss of a crucial proportion of Earth's fauna during the next 50–100 years would be irreparable over the time frame of human existence, and future societies may struggle to live sustainably if left with unstable, depauperate life-support systems.

Our model has some parallels with existing research on traps and transformations<sup>94–97</sup>, with many relevant details and implications that will take time to work out. It is not directly diagnostic or prescriptive but it has the potential to both explain and predict, in the context of ESS and agriculture, the creation and resolution of scale mismatches<sup>98</sup> and the development of various systemic syndromes (such as the retention of perverse incentives and subsidies<sup>99</sup>, or the continued presence of destructive feedback loops that are almost impossible to break). The transition from green to red loops may also help to explain collapses in some past societies, providing a translation mode (in moving from theory to empirical, testable hypotheses) for ideas about social complexity and adaptive cycles<sup>100</sup>. A diversity of data-intensive, comparative case studies is needed to test and refine these ideas; developing economies and fast-growing cities should be particularly fertile grounds for further research. Ultimately, we see in these ideas the basis for a scientific framework that would explain why humanity's use of ESS is, despite our combined knowledge and expertise, rapidly approaching planetary boundaries. ■

Received 11 March; accepted 28 July 2014.

1. Rockström, J. *et al.* A safe operating space for humanity. *Nature* **461**, 472–475 (2009).

**In this paper, the authors report that Earth's resources are finite and we are already living unsustainably in some areas.**

2. Millennium Assessment. *Ecosystems and Human Wellbeing: Biodiversity Synthesis* (Island, 2005).
3. United Nations Department of Economic and Social Affairs Population Division. *World Population Prospects: The 2012 Revision, Volume I: Comprehensive Tables* <http://esa.un.org/wpp/documentation/publications.htm> (United Nations, 2013).
4. Rudel, T. K. *et al.* Agricultural intensification and changes in cultivated areas, 1970–2005. *Proc. Natl Acad. Sci. USA* **106**, 20675–20680 (2009).
5. Foley, J. A. *et al.* Global consequences of land use. *Science* **309**, 570–574 (2005).
6. O'Brien, K. L. & Leichenko, R. M. Double exposure: assessing the impacts of climate change within the context of economic globalization. *Glob. Environ. Change* **10**, 221–232 (2000).
7. Khoury, C. K. *et al.* Increasing homogeneity in global food supplies and the implications for food security. *Proc. Natl Acad. Sci. USA* **111**, 4001–4006 (2014).
8. Rabotyagov, S. S., Kling, C. L., Gassman, P. W., Rabalais, N. N. & Turner, R. E. The economics of dead zones: causes, impacts, policy challenges, and a model of the gulf of Mexico hypoxic zone. *Rev. Environ. Econ. Policy* **8**, 58–79 (2014).
9. Cramer, W. *et al.* Tropical forests and the global carbon cycle: impacts of atmospheric carbon dioxide, climate change and rate of deforestation. *Phil. Trans. R. Soc. Lond. B* **359**, 331–343 (2004).

10. Child, M. F., Cumming, G. S. & Amano, T. Assessing the broad-scale impact of agriculturally transformed and protected area landscapes on avian taxonomic and functional richness. *Biol. Conserv.* **142**, 2593–2601 (2009).
11. Norberg, J. & Cumming, G. S. *Complexity Theory for a Sustainable Future* (Columbia Univ. Press, 2008).
12. Loreau, M. *et al.* Biodiversity and ecosystem functioning: current knowledge and future challenges. *Science* **294**, 804–808 (2001).
13. Tscharntke, T. *et al.* Global food security, biodiversity conservation and the future of agricultural intensification. *Biol. Conserv.* **151**, 53–59 (2012).
14. Millennium Assessment. *Ecosystems and Human Well-being: a Framework for Assessment. A Report of the Conceptual Framework Working Group of the Millennium Ecosystem Assessment* (Island, 2003).
15. Crossman, N. D. *et al.* A blueprint for mapping and modelling ecosystem services. *Ecosyst. Serv.* **4**, 4–14 (2013).
16. Burkhard, B., Kroll, F., Nedkov, S. & Müller, F. Mapping ecosystem service supply, demand and budgets. *Ecol. Indic.* **21**, 17–29 (2012).
17. Martínez-Harms, M. J. & Balvanera, P. Methods for mapping ecosystem service supply: a review. *Inter. J. Biodiv. Sci. Ecosyst. Serv. Mgmt* **8**, 17–25 (2012).
18. Geist, H. J. & Lambin, E. F. Proximate causes and underlying driving forces of tropical deforestation. *Bioscience* **52**, 143–150 (2002).
19. Revilla, E. & Sáenz, M. J. Supply chain disruption management: global convergence vs national specificity. *J. Bus. Res.* **67**, 1123–1135 (2014).
20. Mooney, H. A., Duraipah, A. & Larigauderie, A. Evolution of natural and social science interactions in global change research programs. *Proc. Natl Acad. Sci. USA* **110**, 3665–3672 (2013).
21. Carpenter, S. R. *et al.* Science for managing ecosystem services: beyond the Millennium Ecosystem Assessment. *Proc. Natl Acad. Sci. USA* **106**, 1305–1312 (2009).
22. Meyers, B. *et al.* Getting the measure of ecosystem services: a social-ecological approach. *Front. Ecol. Environ.* **11**, 268–273 (2013).
- This paper reports that policy-related indicators for development goals have focused almost entirely on ecosystems, without effective monitoring of the socioeconomic systems that often drive ecosystem change.**
23. Perrings, C. *et al.* Ecosystem services for 2020. *Science* **330**, 323–324 (2010).
24. Folke, C. *et al.* Regime shifts, resilience, and biodiversity in ecosystem management. *Annu. Rev. Ecol. Syst.* **35**, 557–581 (2004).
25. Scheffer, M. & Westley, F. R. The evolutionary basis of rigidity: locks in cells, minds, and society. *Ecol. Soc.* **12**, 36 (2007).
26. Chapin, F. S. *et al.* Directional changes in ecological communities and social-ecological systems: a framework for prediction based on Alaskan examples. *Am. Nat.* **168**, S36–S49 (2006).
27. Boserup, E. *Population and Technological Change: a Study of Long-term Trends* (Univ. Chicago, 1981).
28. Zeder, M. A. Domestication and early agriculture in the Mediterranean basin: origins, diffusion, and impact. *Proc. Natl Acad. Sci. USA* **105**, 11597–11604 (2008).
29. Livi-Bacci, M. *A Concise History of World Population* (Wiley, 2012).
30. Deutscher Bauernverband (German Farmers' Union). *Situationsbericht 2013* (Deutscher Bauernverband, 2013).
31. Rodriguez, J. P. *et al.* in *Millennium Ecosystem Assessment Volume 2: Scenarios Assessment* Ch. 11 (Island Press, 2005).
32. Raudsepp-Hearne, C., Peterson, G. D. & Bennett, E. M. Ecosystem service bundles for analyzing tradeoffs in diverse landscapes. *Proc. Natl Acad. Sci. USA* **107**, 5242–5247 (2010).
33. Nelson, E. *et al.* Modeling multiple ecosystem services, biodiversity conservation, commodity production, and tradeoffs at landscape scales. *Front. Ecol. Environ.* **7**, 4–11 (2009).
34. Bennett, E. M., Peterson, G. D. & Gordon, L. J. Understanding relationships among multiple ecosystem services. *Ecol. Lett.* **12**, 1394–1404 (2009).
35. Clough, Y. *et al.* Combining high biodiversity with high yields in tropical agroforests. *Proc. Natl Acad. Sci. USA* **108**, 8311–8316 (2011).
36. Popp, A. *et al.* Land-use transition for bioenergy and climate stabilization: model comparison of drivers, impacts and interactions with other land use based mitigation options. *Clim. Change* **123**, 495–509 (2013).
37. Nevens, F., Frantzeskaki, N., Gorissen, L. & Loorbach, D. Urban transition labs: co-creating transformative action for sustainable cities. *J. Cleaner Prod.* **50**, 111–122 (2013).
38. Berkes, F. *Sacred Ecology: Traditional Ecological Knowledge and Resource Management* (Taylor and Francis, 1999).
39. Tengö, M. *Management Practices for Dealing with Uncertainty and Change: Social-Ecological Systems in Tanzania and Madagascar*. PhD thesis, Stockholm Univ. (2004).
40. Soma, T. Contemporary falconry in the Altai-Kazakh in Western Mongolia. *Int. J. Intangible Heritage* **7**, 103–111 (2012).
41. Young, H. & Jacobsen, K. No way back? Adaptation and urbanization of IDP livelihoods in the Darfur Region of Sudan. *Dev. Change* **44**, 125–145 (2013).
42. Smith, A. & Garnier, M. *An Inquiry into the Nature and Causes of the Wealth of Nations* (Nelson, 1845).
43. Romer, P. M. Endogenous technological change. *J. Polit. Econ.* **98**, S71–S102 (1990).
- This paper develops a model that explains how economic growth arises from endogenous technological change.**
44. Matsuyama, K. Agricultural productivity, comparative advantage, and economic growth. *J. Econ. Theory* **58**, 317–334 (1992).
- The key theoretical contribution of this paper is the finding that the effect of agricultural productivity on economic growth and industrialization depends on the openness of an economy.**



45. Wu, J. Urban ecology and sustainability: the state-of-the-science and future directions. *Landscape Urban Plan.* **125**, 209–221 (2014).
46. Luck, M. A., Jenerette, G. D., Wu, J. & Grimm, N. B. The urban funnel model and the spatially heterogeneous ecological footprint. *Ecosystems* **4**, 782–796 (2001).
47. Grimm, N. B. *et al.* Global change and the ecology of cities. *Science* **319**, 756–760 (2008).
48. Ostrom, E. How types of goods and property rights jointly affect collective action. *J. Theor. Polit.* **15**, 239–270 (2003).
- This paper lays out a set of premises that explain the conditions that determine the effectiveness of institutions in common property systems.**
49. McGinnis, M. D. An introduction to IAD and the language of the Ostrom workshop: a simple guide to a complex framework. *Policy Stud. J.* **39**, 169–183 (2011).
50. Scheffer, M. *et al.* Early-warning signals for critical transitions. *Nature* **461**, 53–59 (2009).
51. Carpenter, S. R. & Turner, M. G. Hares and tortoises: interactions of fast and slow variables in ecosystems. *Ecosystems* **3**, 495–497 (2000).
52. Leamer, E. E., Maul, H., Rodriguez, S. & Schott, P. K. Does natural resource abundance increase Latin American income inequality? *J. Dev. Econ.* **59**, 3–42 (1999).
53. Lebel, L. *et al.* Industrial transformation and shrimp aquaculture in Thailand and Vietnam: pathways to ecological, social, and economic sustainability? *AMBIO* **31**, 311–323 (2002).
54. Ocampo, J. & Parra-Lancourt, M. The term of trade for commodity since the mid 19th century. *J. Iberian Latin Am. Econ. Hist.* **28**, 11–43 (2009).
55. DeFries, R. & Pandey, D. Urbanization, the energy ladder and forest transitions in India's emerging economy. *Land Use Policy* **27**, 130–138 (2010).
56. Ishii, H. T. *et al.* Integrating ecological and cultural values toward conservation and utilization of shrine/temple forests as urban green space in Japanese cities. *Landscape Ecol. Eng.* **6**, 307–315 (2010).
57. Genske, D. & Ruff, A. in *Proc. 10th IAEG Int. Congress* 82 <http://www.iaeg.info/iaeg2006/start.html> (2006).
58. Getter, K. L. & Rowe, D. B. The role of extensive green roofs in sustainable development. *HortScience* **41**, 1276–1285 (2006).
59. Hofsten, E. & Lundstrom, H. *Swedish Population History: Main Trends from 1750 to 1970* (National Central Bureau of Statistics, 1976).
60. Lobell, H., Schön, L. & Krantz, O. Swedish historical national accounts, 1800–2000: principles and implications of a new generation. *Scand. Econ. Hist. Rev.* **56**, 142–159 (2008).
61. Einhorn, E. & Logue, J. *Modern Welfare States: Politics and Policies in Social Democratic Scandinavia* (Praeger Publishers, 1989).
62. Schön, L. Internal and external factors in Swedish industrialization. *Scand. Econ. Hist. Rev.* **45**, 209–223 (1997).
63. Björklund, J., Limburg, K. E. & Rydberg, T. Impact of production intensity on the ability of the agricultural landscape to generate ecosystem services: an example from Sweden. *Ecol. Econom.* **29**, 269–291 (1999).
64. Statistics Sweden. <http://www.scb.se/en> (Statistics Sweden, 2013).
65. Krauss, J. *et al.* Habitat fragmentation causes immediate and time-delayed biodiversity loss at different trophic levels. *Ecol. Lett.* **13**, 597–605 (2010).
66. Angelstam, P. *et al.* Protecting forest areas for biodiversity in Sweden 1991–2010: policy implementation process and outcomes on the ground. *Silva Fennica* **45**, 1111–1133 (2011).
67. Holmlund, C. M. & Hammer, M. Ecosystem services generated by fish populations. *Ecol. Econ.* **29**, 253–268 (1999).
68. Jansson, Å., Folke, C., Rockström, J., Gordon, L. & Falkenmark, M. Linking freshwater flows and ecosystem services appropriated by people: the case of the Baltic Sea drainage basin. *Ecosystems* **2**, 351–366 (1999).
69. Stoate, C. *et al.* Ecological impacts of early 21st century agricultural change in Europe — a review. *J. Environ. Manage.* **91**, 22–46 (2009).
70. McIntosh, R. J. & McIntosh, S. K. The inland Niger delta before the empire of Mali: evidence from Jenne-jeno. *J. Afr. Hist.* **22**, 1–22 (1981).
71. Batiano, A., Lompo, F. & Koala, S. Research on nutrient flows and balances in West Africa: state-of-the art. *Agric. Ecosyst. Environ.* **71**, 19–35 (1998).
72. FAOSTAT. *Crops: Primary Production and Trade Databases* <http://faostat.fao.org/> Accessed Dec 2013 (FAO, 2013).
73. Brinkmann, K., Schumacher, J., Dittich, A., Kadaore, I. & Buerkert, A. Analysis of landscape transformation processes in and around four West African cities over the last 50 years. *Landscape Urban Plan.* **105**, 94–105 (2012).
74. Buerkert, A. & Schlecht, E. Agricultural innovations in small-scale farming systems of Sudano-Sahelian West Africa: some prerequisites for success. *Secheresse* **24**, 322–329 (2013).
- This paper reports that factors driving the success of agricultural innovations in sub-Saharan Africa are their capacity to enhance farmers' access to markets, the possibility to adopt an innovation with only small amounts of capital, and limited risk of failure despite high rainfall variability.**
75. Mortimore, M. & Turner, B. Does the Sahelian smallholders' management of woodland, farm trees, rangeland support the hypothesis of human-induced desertification? *J. Arid Environ.* **63**, 567–595 (2005).
76. Tappan, G. & McGahuey, M. Tracking environmental dynamics and agricultural intensification in southern Mali. *Agric. Syst.* **94**, 38–51 (2007).
77. The National Bureau of Statistics of Beijing. *Beijing Statistical Yearbook* (ed. Xiuqin, Y.) (Chinese Statistics, 2010).
78. Zhang, S., Deng, L., Yue, P. & Cui, H. *Study on Water Tariff Reform and Income Impacts in China's Metropolitan Areas: the Case of Beijing* <http://documents.worldbank.org/curated/en/2007/07/10119647/study-water-tariff-reform-income-impacts-chinas-metropolitan-areas-case-beijing> (World Bank, 2007).
79. Huang, J., Zhang, H.-L., Tong, W.-J. & Chen, F. The impact of local crops consumption on the water resources in Beijing. *J. Cleaner Production* **21**, 45–50 (2012).
80. State Environmental Protection Administration of China. *SEPA Report* [in Chinese] <http://www.sepa.gov.cn/eic/652466692596695040/20040602/1050958.shtml> (SEPA, 2004).
81. Johnson, T. M., Liu, F. & Newfarmer, R. *Clear Water, Blue Skies: China's Environment in the New Century* (World Bank, 1997).
82. Liu, X. J. *et al.* Enhanced nitrogen deposition over China. *Nature* **494**, 459–462 (2013).
83. Shen, J. L. *et al.* High concentrations and dry deposition of reactive N species at two sites in the North China Plain. *Environ. Pollut.* **157**, 3106–3113 (2009).
84. Guo, J. H. *et al.* Significant soil acidification in major Chinese croplands. *Science* **327**, 1008–1010 (2010).
85. Zhuang, G. S., Guo, J. H., Yuan, H. & Zhao, C. Y. The compositions, sources, and size distribution of the dust storm from China in spring of 2000 and its impact on the global environment. *Chin. Sci. Bull.* **46**, 895–900 (2001).
86. Gibbs, H. *et al.* Tropical forests were the primary sources of new agricultural land in the 1980s and 1990s. *Proc. Natl Acad. Sci. USA* **107**, 16732–16737 (2010).
87. Sanderson, F. J., Donald, P. F., Pain, D. J., Burfield, I. J. & Van Bommel, F. P. Long-term population declines in Afro-Palearctic migrant birds. *Biol. Conserv.* **131**, 93–105 (2006).
88. Zimmerman, B., Peres, C. A., Malcolm, J. R. & Turner, T. Conservation and development alliances with the Kayapo of south-eastern Amazonia, a tropical forest indigenous people. *Environ. Conserv.* **28**, 10–22 (2001).
89. Smith, E. A. & Wishnie, M. Conservation and subsistence in small-scale societies. *Annu. Rev. Anthropol.* **29**, 493–524 (2000).
90. Daily, G. C. & Ehrlich, P. R. Socioeconomic equity, sustainability, and Earth's carrying capacity. *Ecol. Appl.* **6**, 991–1001 (1996).
- This paper reports that increasing equity can help to increase food production and reduce fertility rates, but runaway consumption must be curbed for sustainability.**
91. Kuznets, S. Economic growth and income inequality. *Am. Econ. Rev.* **45**, 1–28 (1955).
92. Dasgupta, S., Laplante, B., Wang, H. & Wheeler, D. Confronting the environmental Kuznets curve. *J. Econ. Perspect.* **16**, 147–168 (2002).
93. Stern, D. I. The rise and fall of the environmental Kuznets curve. *World Dev.* **32**, 1419–1439 (2004).
94. Costanza, R. L. *et al.* Sustainability or collapse: what can we learn from integrating the history of humans and the rest of nature? *Ambio* **36**, 522–527 (2007).
95. Haug, G. H. *et al.* Climate and the collapse of Maya civilization. *Science* **299**, 1731–1735 (2003).
96. Rasmussen, L. V. & Reenberg, A. Collapse and recovery in Sahelian agro-pastoral systems: rethinking trajectories of change. *Ecol. Soc.* **17**, 14 (2012).
97. Tainter, J. A. *The Collapse of Complex Societies* (Cambridge Univ. Press, 1988).
98. Cumming, G. S., Cumming, D. H. M. & Redman, C. L. Scale mismatches in social-ecological systems: causes, consequences, and solutions. *Ecol. Soc.* **11**, 14 (2006).
- This paper argues that mismatches in the scales at which ecosystems vary and the institutional levels at which responsibility for ecosystem management resides can lead to various management problems.**
99. van Beers, C. & van den Bergh, J. C. Perseverance of perverse subsidies and their impact on trade and environment. *Ecol. Econ.* **36**, 475–486 (2001).
100. Holling, C. S. & Meffe, G. K. Command and control and the pathology of natural resource management. *Conserv. Biol.* **10**, 328–337 (1996).
- This paper states that attempts to maximize off-take from production systems often create vulnerabilities in those systems, leading to collapse.**

**Acknowledgements** G.S.C. thanks the Universität Kassel (<http://www.icdd.uni-kassel.de>) and Georg-August-Universität Göttingen for travel funding. This research was partially supported by a James S. McDonnell Foundation grant to G.S.C. and benefitted from discussions between A.B. and E.S. in BU1308/5-3, SCHL587/4-3 and the UrbanFood project, funded by the Volkswagen Foundation (No. I/82 189); between S.v.C.-T. and T.T. within the RTG 1644 (*Scaling Problems in Statistics*); and the discussions of T.T. within the CRC 990 (EFForTS).

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of this paper at [go.nature.com/t4fvnf](http://go.nature.com/t4fvnf). Correspondence should be addressed to G.S.C. ([gscumming@gmail.com](mailto:gscumming@gmail.com)).

# Learning to coexist with wildfire

Max A. Moritz<sup>1</sup>, Enric Batllori<sup>1†</sup>, Ross A. Bradstock<sup>2</sup>, A. Malcolm Gill<sup>3</sup>, John Handmer<sup>4</sup>, Paul F. Hessburg<sup>5</sup>, Justin Leonard<sup>6</sup>, Sarah McCaffrey<sup>5</sup>, Dennis C. Odion<sup>7</sup>, Tania Schoennagel<sup>8</sup> & Alexandra D. Syphard<sup>9</sup>

**The impacts of escalating wildfire in many regions — the lives and homes lost, the expense of suppression and the damage to ecosystem services — necessitate a more sustainable coexistence with wildfire. Climate change and continued development on fire-prone landscapes will only compound current problems. Emerging strategies for managing ecosystems and mitigating risks to human communities provide some hope, although greater recognition of their inherent variation and links is crucial. Without a more integrated framework, fire will never operate as a natural ecosystem process, and the impact on society will continue to grow. A more coordinated approach to risk management and land-use planning in these coupled systems is needed.**

Fire is unique among the natural hazards that affect human communities and the ecosystems on which we depend<sup>1</sup>. Although humans sometimes intentionally ignite and manage fires, our main focus is on fighting them. For other natural hazards, such as earthquakes, hurricanes and floods, there is much more emphasis on identifying vulnerabilities and adaptations. The ‘command and control’ approach<sup>2</sup> typically used in fire management neglects the fundamental role that fire regimes have in sustaining biodiversity and key ecosystem services<sup>3–6</sup>. Unless people view and plan for fire as an inevitable and natural process, it will continue to have serious consequences for both social and ecological systems.

Over the past two decades, wildfires around the world have increasingly affected human values (for example, lives, views or sacred environments) and assets (for example, damage to homes or public infrastructure) and ecosystem services (for example, air quality and long-term carbon storage). The growing list of negative outcomes and their financial effects have complex causes and consequences<sup>7</sup>. The natural range of fire sizes and resultant frequencies, timings and intensities — the ‘fire regime’ — varies greatly among ecosystems, as do the ways in which human activities have altered them (for example, through timber harvesting, fire suppression, urban or agricultural encroachment, novel ignition patterns and invasive species). Not surprisingly, policy strategies to address wildfires often emphasize fuel reduction<sup>8,9</sup>. However, even where strategies recognize interacting cultural, environmental and economic dimensions of wildfire<sup>10–12</sup>, few tackle the difficult land-use issue of where and how humans choose to build their communities in the first place. The prospect of widely increasing fire activity with climate change<sup>13</sup> intensifies the need for a new path forward.

Viewing fire-related problems in the context of coupled socioecological systems (SESs)<sup>14</sup>, which explicitly recognize links between humans and their natural environments, provides insights into achieving a more sustainable coexistence with wildfire. We have learned a great deal about fire as an essential ecosystem process and the human dimensions of living on fire-prone landscapes. Synthesis of this knowledge through a coupled systems approach can highlight specific vulnerabilities and trade-offs, and facilitate adaptation strategies across widely varying public and private

landscapes (Fig. 1). In this Review, we summarize research on fire-prone ecosystems and fire effects on human communities through the lens of SESs, identify links in these coupled systems, and discuss recommendations for greater resilience. We emphasize insights from three regions (Fig. 2) where major fire-related losses have occurred in recent decades: the Mediterranean basin, the western United States and Australia.

## Socioecological systems and fire

Sustainable solutions to most environmental problems will be impossible if the links and interdependencies between humans and ecosystems are ignored<sup>14</sup>. In the context of wildfire, the most well-developed SES research that incorporates this coupling concerns climate-change effects on Alaskan boreal forest ecosystems and rural indigenous communities<sup>15,16</sup>. Case studies in rural communities of New Zealand<sup>17</sup> and California<sup>18</sup> also exist. Remarkably, a coupled wildfire SES framework has yet to be adopted for the more densely developed wildland–urban interface (WUI; area in which communities intermix with or abut natural vegetation), where most of the human fatalities, home losses and fire-suppression expenditures occur.

The complexity of how wildfire operates in different ecosystems and how humans interact with it indicates that place-based hazards and risks should be addressed as a coupled SES<sup>16,19</sup>. Reframing the problem to minimize harmful effects as the climate changes and humans increasingly inhabit fire-prone landscapes identifies an integrated set of coupled SES linkages (Fig. 1). Importantly, this allows us to recognize how the geographic context of the coupling itself contributes to impacts and losses of assets throughout the wildfire SES. Local characteristics of the WUI, and the components on either side of it, will largely determine the degree to which fire may be accommodated and how communities will be affected. The spatial scale of the coupling may also be broad in some cases, such as when fires compromise recreation values (for example, trail access, camping facilities or fishing habitat) and water supplies of distant urbanized areas, or when concerns over human exposure to drifting smoke influence management decisions about fires that are burning relatively far away. Although this framing does not intrinsically address connections between fire and global-scale climate change mitigation<sup>13,15,20</sup>, it helps to

<sup>1</sup>Department of Environmental Science, Policy, and Management, Division of Ecosystem Sciences, University of California, Berkeley, 130 Mulford Hall, Berkeley, California 94720, USA. <sup>2</sup>University of Wollongong, Northfields Avenue, Wollongong, New South Wales 2522, Australia. <sup>3</sup>Australian National University, Canberra, Australian Capital Territory 0200, Australia. <sup>4</sup>RMIT University, 124 Little La Trobe Street, Melbourne, Victoria 3000, Australia. <sup>5</sup>US Forest Service, 1400 Independence Avenue, SW Washington DC 20250-1111, USA. <sup>6</sup>CSIRO, Clayton South, Victoria 3169, Australia. <sup>7</sup>University of California, Santa Barbara, Santa Barbara, California 93106, USA. <sup>8</sup>University of Colorado, Boulder, Boulder 80309-0450, Colorado, USA. <sup>9</sup>Conservation Biology Institute, 136 SW Washington Avenue, Suite 202, Corvallis, Oregon 97333, USA. <sup>†</sup>Present address: Forest Sciences Center of Catalonia & Center for Ecological Research and Forestry Applications, Pujada del Seminari, 28250 Solsona, Spain.



reveal geographically relevant solutions for decreasing harmful effects and increasing the positive benefits of fire on the landscape. The institutional complexity that underlies many aspects of this coupled SES framework — agency mandates, property rights, building ordinances, indigenous governance, economic subsidies and political pressures — will also feed into a particular set of solutions, often creating challenging constraints.

Sustainable coexistence with wildfire is both a process and a long-term goal, such that policy, planning and management are adapted and refined through time (Fig. 1). Responsibility must be shared between governments and the people at risk, and the approach integrates building, planning, fuel management, suppression capability, and knowledge of fire and ecosystem dynamics at different scales. Coexistence with wildfire should ultimately allow ecologically appropriate fire regimes to operate on landscapes near and far from the WUI, with relatively low risks to people, property and resources, while also allowing us to enjoy ecosystem services enhanced by fire (for example, habitat maintenance, potential hazard reduction, natural hydrologic functioning, and carbon and nutrient cycling). This outcome should also reduce the costs of fire suppression and the need to put firefighters at risk.

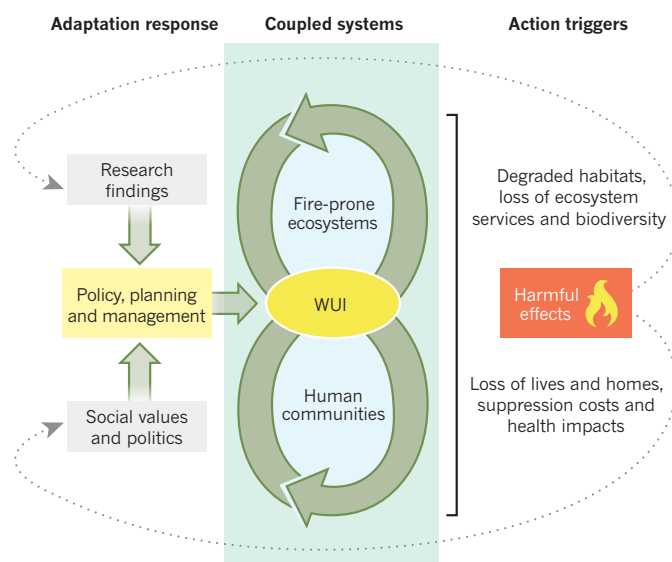
### Fire and ecosystems

The role of fire in different ecosystems varies by the degree of current landscape modification, relative to natural or historical patterns and processes. Some regions have large expanses of semi-wilderness where maintenance or restoration of certain fire regimes is crucial to ongoing habitat characteristics or ecosystem services (for example, the western United States and Australia). Here the links between fire characteristics and ensuing ecological effects, or fire ‘severity’, are often emphasized. Other regions have been so completely altered for various human needs that what is ‘natural’ is no longer a clear consideration (for example, the Mediterranean basin). Furthermore, climatic controls on fire regimes (for example, frequency of droughts or high-wind events, or length of fire season) tend to dominate in some ecosystems, whereas local controls (for example, topography, fuel loads and ignitions) strongly influence others. Fire resilience is thus context-dependent, varying with the biophysical environment and desired future conditions. Accordingly, our capacity to avoid ecosystem degradation and catastrophic shifts<sup>21</sup> (Fig. 1) depends on the ecosystem in question and how climate change will manifest there.

### Mediterranean basin

Mediterranean landscapes are mosaics of various shrublands and oak- and pine-dominated woodlands intermixed with extensive pastures, cultivated lands and abandoned agricultural fields<sup>22</sup>. Despite fire’s ecological influence there<sup>4</sup>, no reference conditions exist for fire management or restoration, and traditional use of fire for rangeland and game management has strongly influenced historical landscape dynamics<sup>23</sup>. Pronounced biophysical and land-use gradients have recently resulted in contrasting fire and vegetation dynamics. The southern and eastern regions are subject to land over-exploitation and reduction in vegetation cover that increases the risk of desertification and loss of ecosystem services. By contrast, socioeconomic drivers are increasing fire hazards and losses over Mediterranean Europe (northern region) owing to rural depopulation, increased WUI exposure and land-cover changes that are sometimes promoted through afforestation policies<sup>24</sup>. Most shrublands and woodlands in the northern region are becoming dense enough to support climate-driven high-intensity ‘crown’ fires<sup>22,25</sup>.

Wildfire in European Union countries is addressed in national and regional forest policy plans, but consensus on fire and ecosystem management is lacking. In spite of large expenditures, increased preparedness and greater firefighting abilities, extreme fire-weather conditions have caused devastating fires in several Mediterranean countries<sup>26</sup>. A new framework to regulate and promote traditional fire practices, accommodating diverse territorial contexts and operational use of fire, has thus been advocated<sup>27</sup>. Currently limited to local management, prescribed burning is increasing across Europe as a tool that aims to reduce fuel loads and diminish the



**Figure 1 | Links and pathways to resilience in coupled socioecological systems affected by fire.** Coexistence with wildfire is strongly influenced by the type of natural fire regimes that operate on a given landscape, and the degree to which communities can reduce exposure and vulnerabilities there. The wildland–urban interface (WUI) is the spatial manifestation of the coupling, and the most proximate scale of exposure and risk mitigation. To learn from and minimize the harmful effects of fire in both the ecosystem and the community, links between systems and scales of interactions must be recognized. Doing so will trigger, through research and in response to changing social values and political context, further adaptation and change in policy, planning and management.

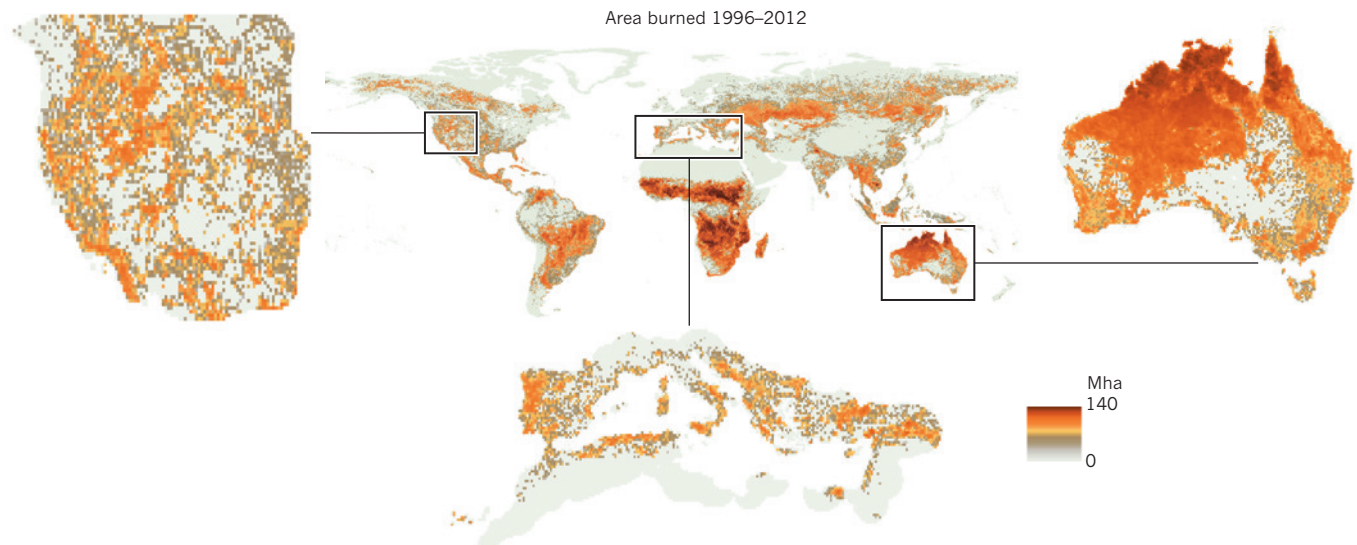
risk of high-intensity fires<sup>28</sup>. Modest changes to regional and national wildfire policies have therefore included long-term preventive actions, but fire management is still primarily centred on short-term fuel- and suppression-oriented measures<sup>8</sup>. There are concerns over the ecological consequences of recent fire patterns<sup>29</sup>, but human-centred fire exclusion generally prevails on most Mediterranean-basin landscapes.

### Western United States

Fire management in many western US ecosystems is informed by research on the historical role of fire<sup>30</sup>, especially through dendrochronology<sup>31</sup> and landscape reconstructions<sup>32</sup>. Before modern management, different types of fire occurred among vegetation types and maintained important natural structures and functions, with great variation geographically<sup>5,32–35</sup>.

In western US forests, high-severity fires that kill overstory trees are typical of cool, high-elevation, subalpine environments<sup>36,37</sup>. Although severe fires may seem catastrophic from a human perspective, in these forests they stimulate vegetation regeneration, promote landscape diversity in terms of vegetation types, provide habitat for many species and sustain other ecosystem services<sup>5</sup>. The many organisms and propagules that may survive the fire, combined with heterogeneity in age, structure and species composition across landscapes, confer resilience against shifts to non-forest types. High-severity fires predominate across about 30% of western US forests, naturally mixing with low-severity fires through time and space across another ~45%<sup>36</sup>. Key regional controls of high-severity fire regimes are extreme drought and high winds<sup>37</sup>, and local (for example, topographic) influences on severity patterns can emerge during less dry conditions<sup>38</sup>. Fuels tend to be naturally abundant in these ecosystems, so modern fire suppression may have decreased historical levels of landscape fragmentation, but it has not increased fuel loads<sup>5,39</sup>.

By contrast, many dry and mesic, low-elevation and mid-montane forests historically experienced more frequent low-severity fires that maintained relatively open forest structures of fire-resistant



**Figure 2 | Area burned patterns and locations of fire-prone regions.** The cumulative area burned between 1996 and 2012 in millions of hectares (Mha) per mapped cell. The western US region consists of the 11 western states in the conterminous United States (left), the Mediterranean basin (middle) contains the Mediterranean-climate biomes and the Australian region (right) encompasses the entire continent (see Supplementary Information).

trees<sup>33,34,40</sup>, across about 25% of western US forests<sup>36</sup>. Ignition patterns, vegetation structure and fuel amount exert a strong control on regimes of frequent low-severity fire, making them more sensitive to modern human perturbations and also more amenable to fuel-management techniques<sup>33,39–41</sup>. Unlike high-severity fire regimes, timber harvesting and decades of fire suppression in drier forests have lengthened intervals, increased densities of smaller trees and shifted regimes of mostly low-severity fires to include more high-severity, stand-replacing fires. The extent to which this has happened is a topic of debate, raising questions about how widespread ‘mixed severity’ fire regimes were prehistorically<sup>32,35,42</sup>. Regardless, reducing accumulated fuels in these forests is often a high management priority. Only where such departures from natural fire regimes have led to denser, multilayered, fire-intolerant forests, however, may fuel-reduction treatments restore more characteristic forest structure and function (Box 1).

There is a general consensus regarding the importance of fire, including the need for prescribed burning, to maintain native grasslands and open woodlands. Woody plant encroachment in many ecosystems with sparse tree cover, driven by a lack of fire and replacement of native herbivores, has reduced plant biodiversity, altered vegetation structure and threatened the fauna that depend on those habitats<sup>43,44</sup>. Fire also plays a crucial part in regeneration for some of the vast shrublands of the western United States, especially California’s densely urbanized chaparral ecosystems. Similar to high-elevation forests, fire in chaparral is stand-replacing and under strong climatic control (patterns of drought and extreme fire weather)<sup>45</sup>, meaning that fuel-reduction efforts have limited effect except in strategic locations<sup>46,47</sup>. Increased fire frequencies, due to abundant human ignitions and non-native grasses that support rapid reburning, threaten to convert many native shrublands to degraded habitats<sup>48</sup>. Invasive grasses also cause very frequent and often large fires across parts of the Great Basin in the western United States<sup>44,49</sup>, driven by the ‘grass-fire cycle’ positive feedback<sup>50</sup> and bringing serious management challenges even to fire-sensitive desert ecosystems<sup>51</sup>.

### Australia

Fire is ubiquitous in Australian ecosystems, including deserts and tropical forests, and a wide range of fire regimes have been mapped using remote sensing<sup>52</sup>. Annual pulses of relatively intense fire dominate the extensive savannahs of northern Australia, with less frequent, massive fires in the

arid zone occurring after above-average rainfall<sup>53</sup>. By contrast, large fires in the temperate forests of the south, although intense, are less extensive and also less regular (decadal occurrence). Biophysical models of fire-regime controls<sup>54</sup> and analysis of trade-offs in fuel characteristics and fire types<sup>52</sup> confirm the primary role of climate, especially the gradient in summer monsoonal precipitation. Thus, fire frequencies tend to vary with latitude, decreasing towards the south and especially the arid interior. Most fire activity on the Australian continent is in grass fuels and of relatively low intensity.

Although palaeo-charcoal deposits document fire’s very long history in Australia<sup>55</sup>, fine-scale understanding of fire-regime variability through dendrochronology is generally lacking, hindering detailed perspectives on long-term variations in fire regimes. Comprehensive fire management initiatives focus on key environmental objectives, such as biodiversity conservation<sup>20</sup> and emissions reduction<sup>56</sup>, as a function of local context. Maintenance of contemporary fire regimes for biodiversity conservation is a priority in most regions, as opposed to the emphasis on restoration that dominates western US approaches.

Australia’s productive eucalyptus forests, which can burn at very high intensities and low–moderate frequencies, are largely restricted to southern and eastern edges of the continent. Although these forests are characteristically Australian, their proximity to urbanized areas has probably fed the continent’s reputation for high-intensity fire events (see ‘Where do people live?’). Debates over the degree to which fuel reduction, whether by mechanical or prescribed fire treatment, can alter the probabilities of high-intensity events<sup>57,58</sup> are similar to those that occur for western US forests.

Prescribed burning in Australia is extensive, but controversial. Fuel reduction burning can partially reduce risk to human life and economic assets, although trade-offs with risks to environmental assets such as biodiversity and ecosystem services are not well understood<sup>3,59</sup>. However, functional responses of species to fire frequencies, sizes, timings and intensities provide a measurable basis for predicting how ecological diversity will respond to management and climate change<sup>60,61</sup>.

### Resilience and climate change

Ecosystem managers in the three regions covered here (Fig. 2) may have limited ability to alter the numbers, sizes and characteristics of fires occurring in different ecosystems<sup>5,34,39,59</sup>. As already discussed, this is because coarse-scale climatic influences tend to control fire regimes in many ecosystems, especially those that are naturally prone to large and high-severity fires. Except under the most extreme conditions, fire regimes typically constrained by more local-scale controls, such as ignition frequencies and biomass accumulation rates, may respond



more strongly to prescribed fire and mechanical fuel reductions. This characterization of two opposing types of fire regimes is, however, a vast over-simplification — idealized end points along a spectrum of variation within and between fire-prone ecosystems<sup>62</sup> — and management prescriptions need to somehow accommodate such complexity. Furthermore, fire-related sensitivities and responses vary among plant and animal species, so fire management for the persistence of one important group of organisms may not favour that of the others.

The potential for climate change to cause ‘novel’ or ‘no analogue’ environmental conditions in some ecosystems presents new challenges for management, policy and planning. An obvious goal is to have ongoing fire regimes that minimize the risk of biodiversity loss<sup>59</sup>. Yet, what adaptation responses are appropriate (Fig. 1) if we do not know how future climates and related biophysical processes will differ from the recent past? These uncertainties have resulted in somewhat similar recommendations about fire and ecosystem resilience<sup>63–65</sup>. Heterogeneity in vegetation types, stand structures and successional age classes at all spatial scales and environmental settings is emerging as a strategy for enhancing ecosystem resilience to climate change. This essentially facilitates diverse initial conditions for multiple future ecological trajectories, the most likely and successful of which will not be known for decades. The role of diverse topography in creating microclimate refugia, or ‘holdouts’<sup>66</sup>, as well as in influencing fire sizes and severity characteristics within large fires<sup>38,67</sup>, comprises the physical template for resilience in more mountainous regions. In ecosystems with a recent paucity of burning, fire management that fosters burning under diverse conditions may be useful for achieving this desired heterogeneity and reducing fuel accumulations<sup>41</sup>. Not all fire-generated heterogeneity is ecologically significant, however, so understanding the effects of specific types of ‘pyrodiversity’ is important<sup>68</sup>.

### Where do people live?

The WUI is the most proximate spatial manifestation of the coupling in a wildfire SES (Fig. 1). Understanding and addressing vulnerabilities related to the WUI in fire-prone areas is therefore crucial to long-term solutions. As distances between urbanized areas and those protected from development decrease globally<sup>69</sup>, a growing WUI will expand the scope of coupling in wildfire SESs worldwide. Negative fire effects that were once due to ‘distant’ fires (for example, the impacts of smoke on human health) will be increasingly common, making coexistence with wildfire much more challenging.

The current WUI of the western United States is relatively well characterized, with over 60% expansion since 1970 (ref. 70) and about 70% in private ownership<sup>71</sup>. The WUI in this region also predominantly occurs where fire severities are high<sup>70</sup>. Only 14% of private land in the western US WUI is developed, so substantial increases in human exposure to fire may occur as the remaining portions become populated<sup>72</sup>. Although less well characterized, there is growing awareness of expanding WUI in Mediterranean Europe<sup>24,73,74</sup> and Australia<sup>19,75</sup>.

Global systematic analyses of human settlement in fire-prone environments is important, but lacking<sup>76</sup>. Coarse-scale characterization of how population densities relate to various fire-prone environments (Fig. 3) provides some insight. Although often characterized as a ‘forest fire’ problem, western US patterns indicate that highly fire-prone locations with large numbers of people tend to be associated with sparse or no tree cover (for example, the chaparral shrublands of southern California); locations with both high population densities and denser forests exhibit the least area burned (Fig. 3, left). Australia exhibits greater area burned over a broader range of environments, with intermediate population densities being more fire-prone regardless of the amount of forest cover (Fig. 3, middle). The Mediterranean basin is unique because the greatest area burned coincides with the highest population densities (Fig. 3, right), although this too occurs in locations with relatively low forest cover (for example, abandoned agricultural lands<sup>26</sup>).

Acknowledging the diversity of the fire-prone environments and vegetation types where people live is important, because it has implications for the types of fuel treatments that may or may not work to mitigate fire hazards within or near the WUI, and it could help to guide future resource allocation decisions (for example, among vegetation removal, evacuation planning and home vulnerability retrofits)<sup>77</sup>. Awareness of the institutional and social diversity of different human communities is also important, as we discuss in the next section, because it influences their capacity for preparation and mitigation of hazards such as wildfires<sup>18</sup>.

### Fire and human communities

This section reviews research on how fires affect human communities and is organized by the scale of coupling in a wildfire SES (Fig. 1), ranging from individuals to landscapes. Social science research on wildfire, primarily undertaken in Australia and the United States,

#### BOX 1

## What can ‘thinning’ of fuels achieve?

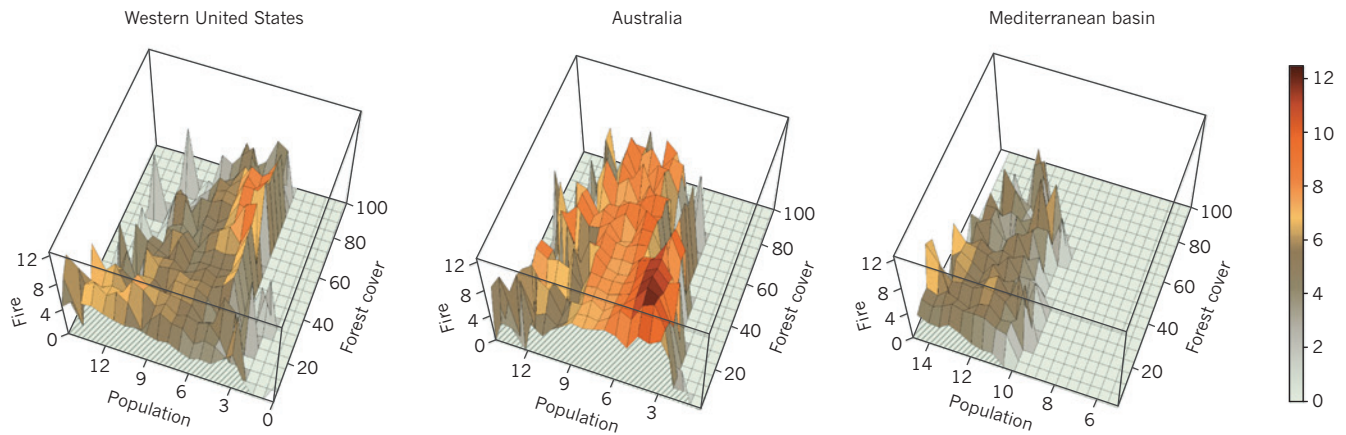
There is intense pressure on land-management agencies to reduce fire hazards (for example, rates of spread or flame lengths if a fire occurs). Treatments should be prioritized, however, where they may help to protect communities or reduce fuel loads in the areas that are most likely to experience uncharacteristically severe burns<sup>36,71</sup>. Mechanical fuel-reduction treatments are most suited to certain dry and fire-prone mesic forests<sup>34,39–41,77</sup>, where thinning the density of smaller understory trees and removing surface fuel residues (non-merchantable tree tops and limbs) created by these treatments can reduce fire intensities and rates of spread<sup>40</sup>. Not treating the additional surface-fuel by-products can actually increase fire intensity and severity when a wildfire does occur<sup>41</sup>.

Some of the most basic trade-offs that limit the widespread use of mechanical fuel reductions involve their economic viability. Often, larger commercial trees will be harvested to help offset operational costs, but this typically generates more surface-fuel residues. Moreover, opening up the overstory canopy and increasing sunlight penetration can increase growth of highly flammable understory

vegetation. Controlling this growth response is an ongoing endeavour, the economic feasibility of which is unknown.

Uncertainty about when and where treatments might actually perform as desired must also be considered. Although there are many examples of fuel treatments reducing fire behaviour when conditions are not extreme, recently treated forests can experience a stand-replacing crown fire when wind speeds exceed 30 km h<sup>-1</sup> and when fuel moisture is low<sup>102</sup>. When the probability of fire occurring in a particular area is relatively low, the odds of a fuel treatment influencing the behaviour of a wildfire there, within the time frame that treatments are effective, is also low<sup>103</sup>. The degree of protection provided by a particular mechanical treatment may thus depend on uncertain parameters (for example, ignition patterns and extreme wind frequencies).

In many areas, ecological restoration and fuel-management goals may be best balanced and accomplished through fire<sup>4,41</sup>, which creates natural heterogeneity and provides for fire-dependent species.



**Figure 3 | Relationship between forest cover, population density and area burned in fire-prone regions.** Locations with both higher human populations and greater amounts of burning tend not to be consistently characterized by high forest cover. Patterns vary greatly among regions, reflecting the different contexts in which each side of the wildfire socioecological system have intersected. (Data were aggregated from

original sources (see Supplementary Information) to 0.25° resolution cells and plotted as density surfaces.) Forest cover is the percentage area covered by trees (>5 m height) per cell in 2000; population is number of people per cell (log transformed) in 2000; and fire is total area burned in hectares per cell (log transformed) between 1996 and 2012. The colour scale for fire is to help differentiate higher peaks in area burned.

is relatively sparse and not easily generalized. Work in the United States emphasizes social acceptance of techniques to mitigate fire risk (for example, fuel reduction on public and private lands) and, more recently, public response during and after fires<sup>78</sup>. In Australia, where many people do not evacuate during fires, risk perception, homeowner preparedness and response during fires, and community safety<sup>79</sup> are key areas of research. We also include studies outside the social sciences that have examined the role of vegetation and fuel treatments linked with losses and the built environment itself.

### Risk perception and public response

Public response to wildfire is shaped by numerous factors, such as local context and individual personality and experience, so simple explanations for action or inaction do not exist. For instance, many researchers and managers assume that individuals do not understand fire risk. But US studies show that most people living in high-fire-risk areas understand their exposure, but there is a tenuous link between understanding risk and taking action to mitigate it; whereas recognizing risk might be necessary to consider mitigation, perceived efficacy of mitigation and resource constraints can be more influential<sup>80</sup>. Similarly, whereas around 80% of people in the fire risk areas of Victoria, Australia, know they are in a hazardous area<sup>81</sup>, this does not necessarily translate to safer actions. After the devastating 2009 Black Saturday fires in Victoria, most people in high-fire-risk areas were aware of what new fire warnings meant and how to ensure their safety, but few acted on the knowledge when the highest-level warning was issued<sup>81</sup>. A deeper understanding of the influences on preparedness, evacuation decisions and support for hazard mitigation is needed.

Specific cultural and institutional systems affect public response to wildfire, as do psychological and social dynamics. For example, institutional structures in the United States and Australia are quite different, but key social dynamics have many similarities. In both countries, trust is a key factor shaping public support for agencies, whether they provide information or engage in fire-management activities<sup>82</sup>. US studies of public acceptance of prescribed fire reveal that trust in the personnel implementing the burn, along with familiarity with the practice, are associated with higher acceptance levels<sup>83</sup>. In terms of the US public response during fires, evacuating has long been the norm, often with mandatory evacuation orders; until Black Saturday, Australians were urged to either prepare to stay and protect their properties, or to leave early, on the basis that either option was safer than leaving late<sup>79</sup>. Despite this difference, the range of public behaviours in both countries is similar, with some residents leaving early, some staying to defend and a substantial number waiting to see how the situation develops. Furthermore, individual actions do not necessarily

reflect a consistent response, as some household members may leave and some stay, while others go back and forth to check on property, animals or those who stay<sup>84</sup>. Although historically 'stay or go' seems to have worked reasonably well in Australia<sup>79</sup>, the approach was questioned after the Black Saturday fires, as it was widely seen to have contributed to many of the 173 deaths. However, roughly half the people (around 3,000 households) in the burnt areas seemed to have stayed and defended their properties successfully and about half left, almost as the fire front was approaching. Most were satisfied with their decision and said they would do the same thing again<sup>84</sup>. Most also stated that they would like to be better prepared. The post-fire effort naturally concentrated on fatalities, with official advice after Black Saturday inquiries shifting to leaving early.

When the public response is to evacuate, key elements to success include environmental conditions (especially fire-weather severity), patterns of roads, neighbourhoods and topography. In Australia, public warnings have been based on a fire-weather danger scale, which was revised after Black Saturday to capture the most extreme conditions, along with altered warning messages and advice for these extremes. There is some public understanding of the reclassification, but little evidence of altered behaviour<sup>81</sup> or understanding that weather conditions well below the extreme level are still dangerous. Analogous fire-weather warnings are issued regularly in other parts of the world, but are not standardized and rarely trigger evacuation orders. Similar to many regions, fatalities during evacuations in the Mediterranean basin tend to occur during the most severe weather conditions, when fires have already begun and people choose to evacuate too late<sup>85</sup>; in addition, such extreme events seem to be on the rise<sup>26</sup>. A growing public safety challenge associated with evacuating people from fire-prone communities in mountainous terrain is limited road access. For example, housing densities are increasing in many WUI regions of the western United States without commensurate increases in the road network to support their evacuation<sup>86</sup>. Emergency planning, including preparation of structures and training for those who choose to stay or simply cannot evacuate safely<sup>87</sup>, is thus increasingly important to the resilience of many communities in the regions reviewed here.

### Structures and surrounding vegetation

To mitigate the risk of structure losses during wildfires, there is increasing evidence from many regions that it is best to focus on the house first and move outward from there<sup>77</sup>. Most structure losses are due to ember attack<sup>88,89</sup>, when flaming or smoldering plant material is lofted by winds and blown inside or against the building or adjacent elements, often long before the flaming front arrives. Embers can cause structure ignition by entering through gaps as small as



2 mm<sup>90</sup> or accumulating outside against flammable building (or surrounding) features. Once ember ignition is addressed through structural design or retrofitting, less prevalent modes of structure loss are important, such as radiant heat and flame exposure. To address these, both building design and surrounding vegetation management are normally considered in unison<sup>19</sup>, with the balance of these treatments being site specific. Similar to evacuation success, an understanding of the local fire-weather conditions and expected types of fires is required<sup>91</sup>. Hence, the building design strategy is to either consider all possible extremes and the weakest link in the system<sup>88</sup> or to pick a threshold level beyond which the structure may not survive. By relating these to a corresponding fire-weather severity, the occupant has the information for deciding when it is necessary to leave early. As a contingency, egress paths from the building interior to another building or area of minimal fuel could improve safety, but preparation for such a fallback is needed long before a wildfire arrives.

Vegetation reduction is most effective immediately adjacent to structures<sup>88,92–94</sup>, as it can eliminate the most immediate sources of combustible material. Vegetation overhanging the structure<sup>91</sup> and ornamental plants<sup>95</sup> have been strongly associated with structure loss. Vegetation clearances more than about 30 m away, however,

seem to provide no significant additional benefit in shrubland environments of southern California, even on steep slopes<sup>94</sup>, reflecting an important trade-off between hazard reduction and habitat values (for organisms dependent on the vegetation removed). Although these findings may only apply to similar shrubland environments, a similar distance to heavily vegetated areas has also been identified for some forested environments, based on radiant heat exposure to structures<sup>77,96</sup>. In Australia, however, a distance from forest edges of more than 30 m was found to influence home losses<sup>93</sup>, indicating that this buffer distance may vary substantially (for example, with fuels, weather and construction types). Another key reason to reduce vegetation near the home is to provide a relatively safe place to engage in structure protection, in case home owners or firefighters are present. It is notable, however, that some species of well-maintained trees (litter removed and high foliar moisture) near the home can actually provide protection, screening embers<sup>19</sup> and acting as a heat sink<sup>96</sup> for an approaching wildfire.

### Landscape-scale patterns

Although fuel treatments seem to provide the greatest protection when located near human communities<sup>19,88,93,94,97</sup>, landscape-scale characteristics of the WUI itself are important. For this reason, a long-term

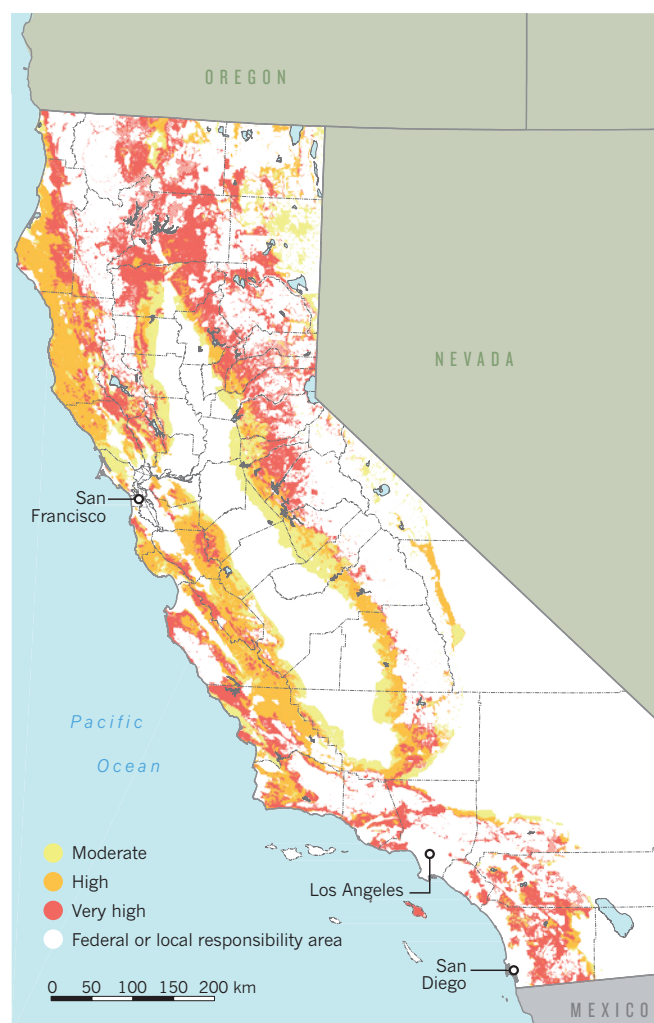
## BOX 2

# Adaptation measures and fire-hazard mapping

Regardless of the surrounding ecosystem conditions, all communities can better coexist with fire by taking several steps: retrofitting homes against ember attack, effectively managing fuels around homes, developing household and community plans for evacuation compared with stay-and-defend decisions, and participating in risk awareness continuing education. For existing high-hazard wildland-urban interface (WUI) areas, landowners may need to take primary responsibility for pursuing the optimal combination of adaptation measures, based on their local vulnerabilities and wildfire exposure. For development of new communities in high-hazard WUI areas, governments need to take a leadership role in planning. Regardless of responsibility, however, all of these efforts will be guided by better mapping of the fire hazard itself.

The fire hazard severity zone (FHSZ) maps (Box Fig.) of California are an official product of the state Department of Forestry and Fire Protection based on a consistent statewide methodology for estimating potential fire behaviour under a set of relatively dry and high wind conditions. Variables that affect modelled fire behaviour include local topography and potential fuel loads, although weather conditions in the current iteration of maps are not tailored to local extremes. Future updates to the FHSZ methodology will incorporate locally varying wind patterns, better reflecting conditions that cause the worst fire-related losses of lives and homes<sup>45,98</sup>.

Fire-resistant residential construction standards are determined by the FHSZ rating of the location in question. In addition, FHSZ classifications must be disclosed at the time of home sales; although this may not deter a sale, it can affect the cost of insuring the home against fire losses. FHSZ maps are thus an incremental but important step towards treating fire like other natural hazards (for example, land-use restrictions associated with flood-plain and earthquake fault maps). Similar mapping methods and codes are produced in Victoria, Australia. Such maps do not explicitly restrict development from occurring — a constraint that should be considered in extremely hazardous locations. Comprehensive approaches should, however, help to better design communities within a complex matrix of both risk and resilience that such maps could reflect spatially. (See Supplementary Information).



approach involving land-use planning offers great potential for reducing wildfire impacts in human communities. A greater understanding is needed concerning building configuration in the WUI and how it relates to risk of losses and fatalities in various environments<sup>73,74</sup>. In some shrubland-dominated landscapes, the arrangement and location of homes have been the most important factors for explaining structure loss: landscape factors such as low housing density, isolated clusters of residential development and long distances to major roads are better predictors of house loss than local factors such as defensible space, fuel or terrain<sup>94,98</sup>. Whether these findings apply to fire-prone landscapes in general or whether there are variations between development patterns and fire regimes needs further research. Although isolated clusters of development and low housing density mean that homes are embedded within, and more exposed to, a matrix of wildland vegetation<sup>19</sup>, ignition-prone homes that are closely spaced in neighbourhoods can also facilitate the spread of house-to-house fire, especially during extreme fire weather.

### Achieving a sustainable coexistence with wildfire

A coupled SES view of wildfire highlights the variation in each half of the SES, as well as how they come together at the WUI, to create many permutations of hazards and vulnerabilities for both human and natural systems. As such, there will be different thresholds for how harmful effects trigger action before, during and after wildfires, and competing societal pressures will influence the degree to which scientific findings are able to guide adaptive responses (Fig. 1). Despite such complexity, some priorities for future work emerge from the extensive research reviewed here.

Context-specific and place-based approaches will be needed to address many existing and future coupled wildfire SES problems. This is because certain fire regimes are inherently more amenable to management activities than others, and also due to the institutional and social diversity that influences human capacity for mitigating risks to individuals and their communities. It is possible, however, that the permutations mentioned above collapse into characteristic typologies that could inform more systematic analyses. If so, are there mutually resilient combinations that are well matched or somehow compatible? Some fire regimes might dictate the degree to which evacuations should be mandatory or how resources might be allocated (for example, training homeowners to protect homes compared with fuel reduction or structure retrofits). A deeper understanding of the variation, links and scales of causes and effects in coupled wildfire SESs is therefore vital.

Governments have a primary responsibility in the long-term evolution of the WUI and the degree to which it limits or amplifies trans-boundary threats in coupled wildfire SESs, so much greater attention to land-use planning is warranted. Land-use regulations to guide fire-related building codes (Box 2) or restrict development in the most fire-prone locations<sup>2,26,99,100</sup> are clearly important steps that government agencies could take to manage the coupling in a wildfire SES. Agencies have a deeper role, however, in the growth of these trans-boundary threats. For example, the 'safe development paradox' applied to flood and hurricane protection demonstrates that making hazardous areas safer for human habitation in the short term actually increases the potential for severe losses over longer time scales<sup>101</sup>. Given that government agencies around the world have focused on reducing fire hazards (for example, through subsidized fire suppression and/or fuel reduction), much less attention has been paid to the ways in which vulnerable WUI development might have been designed from the start. As further development occurs and the WUI expands, so does the need for increased hazard reduction. A perverse consequence of the typical human reaction to fire — to fight it instead of accommodate it — thus contributes to a deepening of coupled wildfire SES problems.

Strategically addressing threats at the WUI maximizes the potential for both effective risk mitigation within developments and management for sustainable fire regimes over the broader sweep of

landscapes. Ultimately, trade-offs and sacrifices must be made to balance these competing demands, but concentration of management effort for risk mitigation in the WUI minimizes the area where adverse effects on environmental assets are likely. Better maps of fire hazards, ecosystem services and climate change effects are thus important for assessing these and other related trade-offs. Addressing all social, economic and environmental assets at risk will necessarily focus on separating those that require exclusion of fire from those where fires of some sort are desirable or inevitable. However, it is unlikely that any planning or management regime will completely exclude fires from vulnerable developments on many landscapes (considerable residual risk to people and property will endure). The capacity for communities to cope with the inevitability of fire, as well as its effects at multiple scales, will therefore be essential.

There is a great deal of research to support better policy, planning and management in all aspects of the coupled wildfire SES problem. Viewing fire as a natural and inevitable hazard should be central to most solutions, so we can anticipate its important positive and negative effects on both human and natural systems. Given that combustion is one of the most basic and ongoing natural processes on Earth, we must continue to learn from our experiences to achieve a sustainable coexistence with wildfire. ■

Received 16 June; accepted 15 September 2014.

- McCaffrey, S. Thinking of wildfire as a natural hazard. *Soc. Nat. Resour.* **17**, 509–516 (2004).  
**This article identifies the importance of viewing fire in the context of natural hazards, which emphasizes human-hazard interactions in ways that most fire research does not.**
- Holling, C. S. & Meffe, G. K. Command and control and the pathology of natural resource management. *Conserv. Biol.* **10**, 328–337 (1996).
- Driscoll, D. A. *et al.* Fire management for biodiversity conservation: key research questions and our capacity to answer them. *Biol. Conserv.* **143**, 1928–1939 (2010).  
**This article examines knowledge gaps and defines a research agenda to better understand species-specific responses to fire regimes, spatio-temporal effects on biota and interactions of fire regimes with other processes.**
- Naveh, Z. in *The Role of Fire in Mediterranean-type Ecosystems* (eds Moreno, J. & Oechel, W.) 163–185 (Springer, 1994).
- Noss, R. F., Franklin, J. F., Baker, W. L., Schoennagel, T. & Moyle, P. B. Managing fire-prone forests in the western United States. *Front. Ecol. Environ.* **4**, 481–487 (2006).  
**This article discusses the science underpinning the development and implementation of fire and fuel management policies for forests before, during and after wildfires.**
- van Wilgen, B. W., Cowling, R. M. & Burgers, C. J. Valuation of ecosystem services. *Bioscience* **46**, 184–189 (1996).
- Gill, A. M., Stephens, S. L. & Cary, G. J. The worldwide "wildfire" problem. *Ecol. Appl.* **23**, 438–454 (2013).  
**This article provides an overview of fire effects on various environmental, social and economic assets, highlighting the complex and geographically specific context of the problem.**
- Fernandes, P. M. Fire-smart management of forest landscapes in the Mediterranean basin under global change. *Landsc. Urban Plan.* **110**, 175–182 (2013).
- Forests and Rangelands. *The National Strategy: The Final Phase in the Development of the National Cohesive Wildland Fire Management Strategy* <http://www.forestsandrangelands.gov/strategy> (US Forests and Rangelands, 2014).
- FAO. *Fire Management: Voluntary Guidelines. Principles and Strategic Actions* <http://www.fao.org/forestry/site/35853/en> (FAO, 2006).
- Forest Fire Management Group. *National Bushfire Management — Policy Statement for Forests and Rangelands* (Forest Fire Management Group for the Council of Australian Governments, 2012).
- Myers, R. L. *Living with Fire-Sustaining Ecosystems and Livelihoods Through Integrated Fire Management* (The Nature Conservancy Global Fire Initiative, 2006).
- Moritz, M. A. *et al.* Climate change and disruptions to global fire activity. *Ecosphere* **3**, 49 (2012).  
**This article provides projections of future fire activity under climate change scenarios and examines sources of uncertainty in such predictions.**
- Berkes, F. & Folke, C. in *Linking Social and Ecological Systems: Management Practices and Social Mechanisms for Building Resilience* (eds Berkes, F. & Folke, C.) 13–20 (Cambridge Univ. Press, 1998).
- Chapin, F. S. *et al.* Policy strategies to address sustainability of Alaskan boreal forests in response to a directionally changing climate. *Proc. Natl Acad. Sci. USA* **103**, 16637–16643 (2006).
- Chapin, F. S. *et al.* Increasing wildfire in Alaska's boreal forest: pathways to



- potential solutions of a wicked problem. *Bioscience* **58**, 531–540 (2008).
17. Jakes, P. J. & Langer, E. R. L. The adaptive capacity of New Zealand communities to wildfire. *Int. J. Wildland Fire* **21**, 764–772 (2012).
  18. Paveglio, T. B., Jakes, P. J., Carroll, M. S. & Williams, D. R. Understanding social complexity within the wildland–urban interface: a new species of human habitation? *Environ. Mgmt* **43**, 1085–1095 (2009).
  19. Gill, A. M. Landscape fires as social disasters: an overview of ‘the bushfire problem’. *Environ. Hazards* **6**, 65–80 (2005).
  20. Bradstock, R. A., Gill, A. M. & Williams, R. J. in *Flammable Australia: Fire Regimes, Biodiversity and Ecosystems in a Changing World* (eds Bradstock, R. A., Gill, A. M. & Williams, R. J.) 307–324 (CSIRO Publishing, 2012).
  21. Scheffer, M., Carpenter, S., Foley, J. A., Folke, C. & Walker, B. Catastrophic shifts in ecosystems. *Nature* **413**, 591–596 (2001).
  22. Mazzoleni, S., di Pasquale, G., Mulligan, M., di Martino, P. & Rego, F. *Recent Dynamics of the Mediterranean Vegetation and Landscape* (Wiley, 2004).
  23. Trabaud, L. V., Christensen, N. L. & Gill, A. M. in *Fire in the Environment: the Ecological, Atmospheric, and Climatic Importance of Vegetation Fires* (eds Crutzen, P. J. & Goldammer, J. G.) 277–295 (Wiley, 1993).
  24. Moreira, F. et al. Landscape–wildfire interactions in southern Europe: implications for landscape management. *J. Environ. Manage.* **92**, 2389–2402 (2011).
  25. Pausas, J. G. & Fernández-Muñoz, S. Fire regime changes in the Western Mediterranean Basin: from fuel-limited to drought-driven fire regime. *Clim. Change* **110**, 215–226 (2012).
  26. San-Miguel-Ayaz, J., Moreno, J. M. & Camia, A. Analysis of large fires in European Mediterranean landscapes: lessons learned and perspectives. *For. Ecol. Manage.* **294**, 11–22 (2013).
  27. Silva, J. S., Rego, F., Fernandes, P. & Rigolot, E. *Towards Integrated Fire Management: Outcomes of the European Project Fire Paradox* (European Forest Institute, 2010).
  28. Castellnou, M., Kraus, D. & Miralles, M. in *Best Practices of Fire Use: Prescribed Burning and Suppression: Fire Programmes in Selected Case-study Regions in Europe* (eds Montiel, C. & Kraus, D. T.) 3–16 (European Forest Institute, 2010).
  29. Koutsias, N. et al. Where did the fires burn in Peloponnisos, Greece the summer of 2007? Evidence for a synergy of fuel and weather. *Agric. For. Meteorol.* **156**, 41–53 (2012).
  30. Swetnam, T. W., Allen, C. D. & Bentacourt, J. L. Applied historical ecology: using the past to manage for the future. *Ecol. Appl.* **9**, 1189–1206 (1999).
  31. Grissino-Mayer, H. D. & Fritts, H. C. The International Tree-Ring Data Bank: an enhanced global database serving the global scientific community. *Holocene* **7**, 235–238 (1997).
  32. Hessburg, P. F., Salter, R. B. & James, K. M. Re-examining fire severity relations in pre-management era mixed conifer forests: inferences from landscape patterns of forest structure. *Landscape Ecol.* **22**, 5–24 (2007).
  33. Allen, C. D. et al. Ecological restoration of southwest ponderosa pine ecosystems: a broad perspective. *Ecol. Appl.* **12**, 1418–1433 (2002).
  34. Keeley, J. E. et al. *Ecological Foundations for Fire Management in North American Forest and Shrubland Ecosystems* (US Department of Agriculture, Forest Service, Pacific Northwest Research Station, 2009).
  35. Perry, D. A. et al. The ecology of mixed severity fire regimes in Washington, Oregon, and Northern California. *For. Ecol. Manage.* **262**, 703–717 (2011).
  36. Schoennagel, T. & Nelson, C. R. Restoration relevance of recent National Fire Plan treatments in forests of the western United States. *Front. Ecol. Environ.* **9**, 271–277 (2011).
  37. Turner, M. G. & Romme, W. H. Landscape dynamics in crown fire ecosystems. *Landscape Ecol.* **9**, 59–77 (1994).
  38. Dillon, G. K. et al. Both topography and climate affected forest and woodland burn severity in two regions of the western US, 1984 to 2006. *Ecosphere* **2**, 130 (2011).
  39. Schoennagel, T., Veblen, T. T. & Romme, W. H. The interaction of fire, fuels, and climate across rocky mountain forests. *Bioscience* **54**, 661–676 (2004).
  40. Agee, J. K. & Skinner, C. N. Basic principles of forest fuel reduction treatments. *For. Ecol. Manage.* **211**, 83–96 (2005).
  41. Stephens, S. L. et al. The effects of forest fuel-reduction treatments in the United States. *Bioscience* **62**, 549–560 (2012).
  42. Odion, D. C. et al. Examining historical and current mixed-severity fire regimes in ponderosa pine and mixed-conifer forests of western North America. *PLoS ONE* **9**, e87852 (2014).
  43. Fuhlendorf, S. D., Woodward, A. J., Leslie, D. M. & Shackford, J. S. Multi-scale effects of habitat loss and fragmentation on lesser prairie-chicken populations of the US Southern Great Plains. *Landscape Ecol.* **17**, 617–628 (2002).
  44. Van Auken, O. W. Causes and consequences of woody plant encroachment into western North American grasslands. *J. Environ. Manage.* **90**, 2931–2942 (2009).
  45. Moritz, M. A., Moody, T. J., Krawchuk, M. A., Hughes, M. & Hall, A. Spatial variation in extreme winds predicts large wildfire locations in chaparral ecosystems. *Geophys. Res. Lett.* <http://dx.doi.org/10.1029/2009GL041735> (2010).
  46. Moritz, M. A., Keeley, J. E., Johnson, E. A. & Schaffner, A. A. Testing a basic assumption of shrubland fire management: how important is fuel age? *Front. Ecol. Environ.* **2**, 67–72 (2004).
  47. Syphard, A. D., Keeley, J. E. & Brennan, T. J. Comparing the role of fuel breaks across southern California national forests. *For. Ecol. Manage.* **261**, 2038–2048 (2011).
  48. Zedler, P. H., Gautier, C. R. & McMaster, G. S. Vegetation change in response to extreme events: the effect of a short interval between fires in California chaparral and coastal scrub. *Ecology* **64**, 809–818 (1983).
  49. Balch, J. K., Bradley, B. A., D’Antonio, C. M. & Gómez-Dans, J. Introduced annual grass increases regional fire activity across the arid western USA (1980–2009). *Glob. Change Biol.* **19**, 173–183 (2013).
  50. D’Antonio, C. M. & Vitousek, P. M. Biological invasions by exotic grasses, the grass/fire cycle, and global change. *Annu. Rev. Ecol. Syst.* **23**, 63–87 (1992).
  51. Brooks, M. L. et al. Effects of invasive alien plants on fire regimes. *Bioscience* **54**, 677–688 (2004).
  52. Murphy, B. P. et al. Fire regimes of Australia: a pyrogeographic model system. *J. Biogeogr.* **40**, 1048–1058 (2013).
  53. Russell-Smith, J. et al. Bushfires ‘down under’: patterns and implications of contemporary Australian landscape burning. *Int. J. Wildland Fire* **16**, 361–377 (2007).
  54. Bradstock, R. A. A biogeographic model of fire regimes in Australia: current and future implications. *Glob. Ecol. Biogeogr.* **19**, 145–158 (2010).
  55. Mooney, S. D., Harrison, S. P., Bartlein, P. & Stevenson, J. in *Flammable Australia: Fire Regimes, Biodiversity and Ecosystems in a Changing World* (eds Bradstock, R. A., Gill, A. M. & Williams, R. J.) 293–305 (CSIRO, 2012).
  56. Cook, G. D., Jackson, S. & Williams, R. J. in *Flammable Australia: Fire Regimes, Biodiversity and Ecosystems in a Changing World* (eds Bradstock, R. A., Gill, A. M. & Williams, R. J.) 293–305 (CSIRO, 2012).
  57. Attiwill, P. M. et al. Timber harvesting does not increase fire risk and severity in wet eucalypt forests of southern Australia. *Conserv. Lett.* **7**, 341–354 (2013).
  58. Price, O. F. & Bradstock, R. A. The efficacy of fuel treatment in mitigating property loss during wildfires: insights from analysis of the severity of the catastrophic fires in 2009 in Victoria, Australia. *J. Environ. Manage.* **113**, 146–157 (2012).
  59. Penman, T. D. et al. Prescribed burning: how can it work to conserve the things we value? *Int. J. Wildland Fire* **20**, 721–733 (2011).
  60. Nimmo, D. G., Kelly, L. T., Farnsworth, L. M., Watson, S. J. & Bennett, A. F. Why do some species have geographically varying responses to fire history? *Ecography* **37**, 805–813 (2014).
  61. Pausas, J. G. & Bradstock, R. A. Fire persistence traits of plants along a productivity and disturbance gradient in mediterranean shrublands of south-east Australia. *Glob. Ecol. Biogeogr.* **16**, 330–340 (2007).
  62. Krawchuk, M. A. & Moritz, M. A. Constraints on global fire activity vary across a resource gradient. *Ecology* **92**, 121–132 (2011).
  63. Dunlop, M. et al. *The Implications of Climate Change for Biodiversity Conservation and the National Reserve System: Final Synthesis* (CSIRO Climate Adaptation Flagship, 2012).
  64. Millar, C. I., Stephenson, N. L. & Stephens, S. L. Climate change and forests of the future: managing in the face of uncertainty. *Ecol. Appl.* **17**, 2145–2151 (2007).
  65. Moritz, M. A., Hurteau, M. D., Suding, K. N. & D’Antonio, C. M. Bounded ranges of variation as a framework for future conservation and fire management. *Ann. NY Acad. Sci.* **1286**, 92–107 (2013).
  66. Hannah, L. et al. Fine-grain modeling of species’ response to climate change: holdouts, stepping-stones, and microrefugia. *Trends Ecol. Evol.* **290**, 390–397 (2014).
  67. Moritz, M. A., Hessburg, P. F. & Povak, N. A. in *The Landscape Ecology of Fire* (eds McKenzie, D., Miller, C., & Falk, D.) 51–86 (Springer, 2011).
  68. Parr, C. L. & Andersen, A. N. Patch mosaic burning for biodiversity conservation: a critique of the pyrodiversity paradigm. *Conserv. Biol.* **20**, 1610–1619 (2006).
  69. McDonald, R. I. et al. Urban effects, distance, and protected areas in an urbanizing world. *Landsc. Urban Plan.* **93**, 63–75 (2009).
  70. Theobald, D. M. & Romme, W. H. Expansion of the US wildland–urban interface. *Landsc. Urban Plan.* **83**, 340–354 (2007).
  71. Schoennagel, T., Nelson, C. R., Theobald, D. M., Carnwath, G. C. & Chapman, T. B. Implementation of National Fire Plan treatments near the wildland–urban interface in the western United States. *Proc. Natl Acad. Sci. USA* **106**, 10706–10711 (2009).
  72. Gude, P., Rasker, R. & van den Noort, J. Potential for future development on fire-prone lands. *J. For.* **106**, 198–205 (2008).
  73. Galiana-Martin, L., Herrero, G. & Solana, J. A wildland–urban interface typology for forest fire risk management in Mediterranean areas. *Landscape Res.* **36**, 151–171 (2011).
  74. Lampin-Maillet, C. et al. Mapping wildland–urban interfaces at large scales integrating housing density and vegetation aggregation for fire prevention in the South of France. *J. Environ. Manage.* **91**, 732–741 (2010).
  75. Lowell, K. et al. Assessing the capabilities of geospatial data to map built structures and evaluate their bushfire threat. *Int. J. Wildland Fire* **18**, 1010–1020 (2009).
  76. Bar-Massada, A., Radeloff, V. C. & Stewart, S. I. Biotic and abiotic effects of human settlements in the wildland–urban interface. *Bioscience* **64**, 429–437 (2014).
  77. Calkin, D. E., Cohen, J. D., Finney, M. A. & Thompson, M. P. How risk management can prevent future wildfire disasters in the wildland–urban interface. *Proc. Natl Acad. Sci. USA* **111**, 746–751 (2014).
  78. McCaffrey, S., Toman, E., Stidham, M. & Shindler, B. Social science research related to wildfire management: an overview of recent findings and future research needs. *Int. J. Wildland Fire* **22**, 15–24 (2013).
  79. Handmer, J. & Tibbitts, A. Is staying at home the safest option during bushfires? Historical evidence for an Australian approach. *Glob. Environ. Change* **6**, 81–91 (2005).

This article discusses the historical basis for Australia’s ‘prepare, stay and defend, or leave early’ policy approach to wildfire.

80. Toman, E., Stidham, M., McCaffrey, S. & Shindler, B. *Social Science at the Wildland-Urban Interface: a Compendium of Research Results to Create Fire-Adapted Communities* (US Department of Agriculture, 2013).
81. Whittaker, J. & Handmer, J. Community bushfire safety: a review of post-black Saturday research. *Aus. J. Emerg. Mgmt* **25**, 7–13 (2010).
82. Olsen, C. S. & Sharp, E. Building community–agency trust in fire-affected communities in Australia and the United States. *Int. J. Wildland Fire* **22**, 822–831 (2013).
83. McCaffrey, S. M. & Olsen, C. S. *Research Perspectives on the Public and Fire Management: a Synthesis of Current Social Science on Eight Essential Questions* (US Department of Agriculture, 2012).
84. Whittaker, J., Haynes, K., Handmer, J. & McLennan, J. Community safety during the 2009 Australian Black Saturday bushfires: an analysis of household preparedness and response. *Int. J. Wildland Fire* **22**, 841–849 (2013).
85. Viegas, D. X., Ribeiro, L., Viegas, M., Pita, L. & Rossa, C. in *Earth Observation of Wildland Fires in Mediterranean Ecosystems* (ed. Chuvieco, E.) 97–109 (Springer, 2009).
86. Cova, T. J., Theobald, D. M., Norman, J. B. & Siebeneck, L. K. Mapping wildfire evacuation vulnerability in the western US: the limits of infrastructure. *GeoJournal* **78**, 273–285 (2013).
87. Penman, T. D. *et al.* Defining adequate means of residents to prepare property for protection from wildfire. *Int. J. Disaster Risk Reduction* **6**, 67–77 (2013).  
**This article discusses the different aspects of what people need to understand to live safely in a fire-prone environment, including the possibility of having to stay and defend a home during a wildfire situation.**
88. Blanchi, R. & Leonard, J. in *Community Bushfire Safety* (eds Handmer, J. & Haynes, K.) 77–85 (CSIRO Publishing, 2008).
89. Cohen, J. Preventing disaster — home ignitability in the wildland-urban interface. *J. For.* **98**, 15–21 (2000).
90. Manzello, S. L., Park, S.-H. & Cleary, T. G. Investigation on the ability of glowing firebrands deposited within crevices to ignite common building materials. *Fire Saf. J.* **44**, 894–900 (2009).
91. Leonard, J. *et al.* *Building and Land-use Planning Research After the 7th February Victorian Bushfires: Preliminary Findings* (CSIRO and Bushfire CRC, 2009).
92. Foote, E. I. D., Martin, R. E. & Gillespie, J. K. in *Proc. 11th Conf. Fire Forest Meteorology* (eds Andrews, P.L. & Potts, D.F.) 16–19 (Society of American Foresters, 1991).
93. Gibbons, P. *et al.* Land management practices associated with house loss in wildfires. *PLoS ONE* **7**, e29212 (2012).  
**This article presents an analyses of factors leading to residential home losses in the 2009 Black Saturday fires in Australia.**
94. Syphard, A. D., Brennan, T. J. & Keeley, J. E. The role of defensible space for residential structure protection during wildfires. *Inter. J. Wildland Fire* <http://dx.doi.org/10.1071/wf13158> (2014).
95. Franklin, S. E. California's catastrophic intermix fires causes, culprits and cures. *Am. Fire J.* **40**, 20–23 (1996).
96. Cohen, J. D. Relating flame radiation to home ignition using modeling and experimental crown fires. *Can. J. For. Res.* **34**, 1616–1626 (2004).
97. Stockmann, K., Burchfield, J., Calkin, D. & Venn, T. Guiding preventative wildland fire mitigation policy and decisions with an economic modeling system. *For. Policy Econ.* **12**, 147–154 (2010).
98. Syphard, A. D., Keeley, J. E., Massada, A. B., Brennan, T. J. & Radeloff, V. C. Housing arrangement and location determine the likelihood of housing loss due to wildfire. *PLoS ONE* **7**, e33954–e33954 (2012).  
**This article presents an analysis of factors leading to residential home losses in fire-prone southern California.**
99. Bovio, G. & Camia, A. Land zoning based on fire history. *Int. J. Wildland Fire* **7**, 249–258 (1997).
100. Buxton, M., Haynes, R., Mercer, D. & Butt, A. Vulnerability to bushfire risk at Melbourne's urban fringe: the failure of regulatory land use planning. *Geogr. Res.* **49**, 1–12 (2011).
101. Burby, R. J. Hurricane Katrina and the paradoxes of government disaster policy: bringing about wise governmental decisions for hazardous areas. *Ann. Am. Acad. Pol. Soc. Sci.* **604**, 171–191 (2006).
102. Cruz, M. G. & Alexander, M. E. Assessing crown fire potential in coniferous forests of western North America: a critique of current approaches and recent simulation studies. *Int. J. Wildland Fire* **19**, 377–398 (2010).
103. Rhodes, J. J. & Baker, W. L. Fire probability, fuel treatment effectiveness and ecological tradeoffs in western U.S. public forests. *Open Forest Sci. J.* **1**, 1–7 (2008).

**Supplementary Information** is linked to the online version of the paper at [www.nature.com/nature](http://www.nature.com/nature).

**Acknowledgements** We would like to thank V. Butsic, S. Cole Moritz, C. English and K. McLeod for comments on drafts of the manuscript, as well as P. Morgan for suggestions that greatly improved the final version. Some of this work was conducted while M.A.M. was a Center Fellow at the National Center for Ecological Analysis and Synthesis, a Center funded by NSF (Grant #EF-0553768), the University of California, Santa Barbara, and the State of California.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of this paper at [go.nature.com/go3vvy](http://go.nature.com/go3vvy). Correspondence should be addressed to M.A.M. ([mmoritz@berkeley.edu](mailto:mmoritz@berkeley.edu)).



# The performance and potential of protected areas

James E. M. Watson<sup>1,2,3</sup>, Nigel Dudley<sup>1,4</sup>, Daniel B. Segan<sup>2,3</sup> & Marc Hockings<sup>1,5</sup>

**Originally conceived to conserve iconic landscapes and wildlife, protected areas are now expected to achieve an increasingly diverse set of conservation, social and economic objectives. The amount of land and sea designated as formally protected has markedly increased over the past century, but there is still a major shortfall in political commitments to enhance the coverage and effectiveness of protected areas. Financial support for protected areas is dwarfed by the benefits that they provide, but these returns depend on effective management. A step change involving increased recognition, funding, planning and enforcement is urgently needed if protected areas are going to fulfil their potential.**

The past few decades have seen protected areas undergo a pronounced expansion, geographically and conceptually. The collective decisions of governments, publicly funded bodies and local communities have created the rapid growth of protected areas throughout the world<sup>1</sup>; land and sea management has seldom changed so quickly over such a large area (Fig. 1). At the same time, as many natural ecosystems fragment<sup>2</sup>, the expectations placed on protected areas by a growing diversity of stakeholders have dramatically increased. Protected areas are now created not only to conserve iconic landscapes and seascapes and to provide habitat for endangered wildlife, but also to contribute to the livelihood of local communities, to bolster national economies through tourism revenues, to replenish fisheries and to play a key part in the mitigation of, and adaptation to, climate change, among many other functions<sup>3</sup>. Importantly, these new demands are in addition to, rather than as a replacement for, earlier motivations, necessitating trade-offs between competing objectives<sup>4,5</sup>. Although the expanded role of protected areas may have fuelled their establishment, their constantly changing focus makes them vulnerable to accusations of failure to achieve one or more of these objectives.

In this Review we explore the extent to which the potential of the global protected area estate is being met and outline the ways in which performance can be improved. We first discuss how broad objectives for protected areas have changed over the past century and how this has corresponded with substantial growth in the global protected area estate. We provide evidence that protected areas — when they are well managed and targeted at threats that they can abate — can deliver a wide range of benefits: protecting magnificent landscapes and seascapes, achieving biodiversity conservation and delivering essential ecosystem services. But we show that the current coverage of the global protected area estate is still well short of meeting core 2020 targets outlined in the Convention on Biological Diversity (CBD)<sup>6</sup>. We explore key issues that are affecting protected area effectiveness and provide recent examples of what seems to be a backslide in commitment by some nations through the defunding, downscaling and delisting of protected areas. Given the key role of protected areas in many social and environmental agendas, the current targets set for conservation and development, and the reduced political commitment in some countries, we conclude this Review with a call for a step change in support for the global protected area estate.

## The protected area movement

Protected areas are not a modern concept. In different forms, they have been around for millennia, whether they are sacred sites guarded by indigenous communities, 'tapu' areas for communal resource use in the Pacific island region or hunting areas set aside to benefit the ruling classes<sup>7</sup>. Over the past century the modern concept of protected areas has been developed and refined, and the global protected area estate has grown extremely rapidly, from a handful of sites at the turn of the twentieth century to more than 162,000 legally designated (statutory) national protected areas, covering more than 28.4 million square kilometres (or 5.6% of Earth's surface)<sup>1</sup> (Fig. 1; see Supplementary Methods for details of protected area calculations). The amount of land now found in statutory protected areas comprises a total area greater than that of South and Central America, and the amount of sea comprises a total area greater than that of the Caribbean Sea, South China Sea, Mediterranean Sea and Bering Sea. These numbers are further augmented by many additional protected areas that are not included in the official tally — those established by local communities, indigenous peoples, private individuals, non-profit trusts, religious groups and even corporations<sup>8</sup> — some of which, such as indigenous territories in the Amazon basin, can themselves be extremely large<sup>9</sup>.

The modern protected-area movement had its nineteenth-century origins in North America, Australia, Europe and South Africa, where protected areas were mainly set up to protect spectacular natural features and wildlife<sup>10</sup>, principally in areas with little potential for economic use<sup>11</sup>. Although many were set up with the dual mandates of landscape and species protection and public use, it was not until around the middle of the twentieth century that tourism inside protected areas accelerated<sup>12</sup>. In some developing countries, the income associated with protected areas is of national significance<sup>13</sup>; in Rwanda, tourism revenue from visits to see mountain gorillas inside Volcanoes National Park is now the country's largest source of foreign exchange, raising US\$200 million annually<sup>14</sup>.

Emerging concern over environmental degradation in the last quarter of the twentieth century influenced motives for establishing protected areas<sup>15</sup>. Recognition of the importance of *in situ* conservation led to a marked expansion of the global terrestrial protected area estate in the 1970s as countries moved to establish protected area networks where species and ecosystems could be conserved from the rapid changes

<sup>1</sup>School of Geography, Planning and Environmental Management, University of Queensland, St Lucia, Queensland 4072, Australia. <sup>2</sup>Wildlife Conservation Society, Global Conservation Program, Bronx, New York 10460, USA. <sup>3</sup>School of Biological Sciences, The University of Queensland, St Lucia, Queensland 4072, Australia. <sup>4</sup>Equilibrium Research, 47 The Quays, Cumberland Road, Spike Island, Bristol BS1 6UQ, UK. <sup>5</sup>UNEP-World Conservation Monitoring Centre, Cambridge CD3 0DL, UK.

taking place elsewhere (Fig. 1). The decision to use protected areas as a core conservation strategy seems to be well justified: recent reviews have concluded that, most of the time, well-managed protected areas reduce rates of habitat loss (the chief threat to biodiversity<sup>16</sup>) in both terrestrial<sup>17,18</sup> and marine<sup>19,20</sup> environments. There is also strong evidence that protected areas maintain species population levels (including threatened species) better than other management approaches<sup>17,21–25</sup>. For example, well-managed marine protected areas (MPAs) have been found to contain more than 5 times the total large fish biomass and 14 times the shark biomass compared with fished areas<sup>20</sup>, and a study of 60 terrestrial protected areas in the tropics found that when they are well managed, there is a positive outcome for biodiversity<sup>26</sup>. This protection

also extends to species that have high financial value and are under intense pressure from well-organized criminal groups<sup>27,28</sup>.

However, because of the rapid growth of protected areas in the latter half of the twentieth century, the contact that local communities had with them increased. At times this growth was in conflict with the needs of local communities and with efforts to address poverty and increase economic development<sup>29,30</sup>. This resulted in widespread criticism of management practices in some protected areas by those concerned with human rights. During this period, management priorities began to shift towards greater recognition of the rights of local communities regarding the governance of the areas in which they live; many protected areas now have management regimes that engage local communities and consciously seek to balance conservation with local livelihoods<sup>31</sup>. In addition, the part that protected areas play in poverty alleviation and fostering economic development in the surrounding communities gained increased attention, and there is now evidence of positive contributions in many regions<sup>4,32,33</sup>.

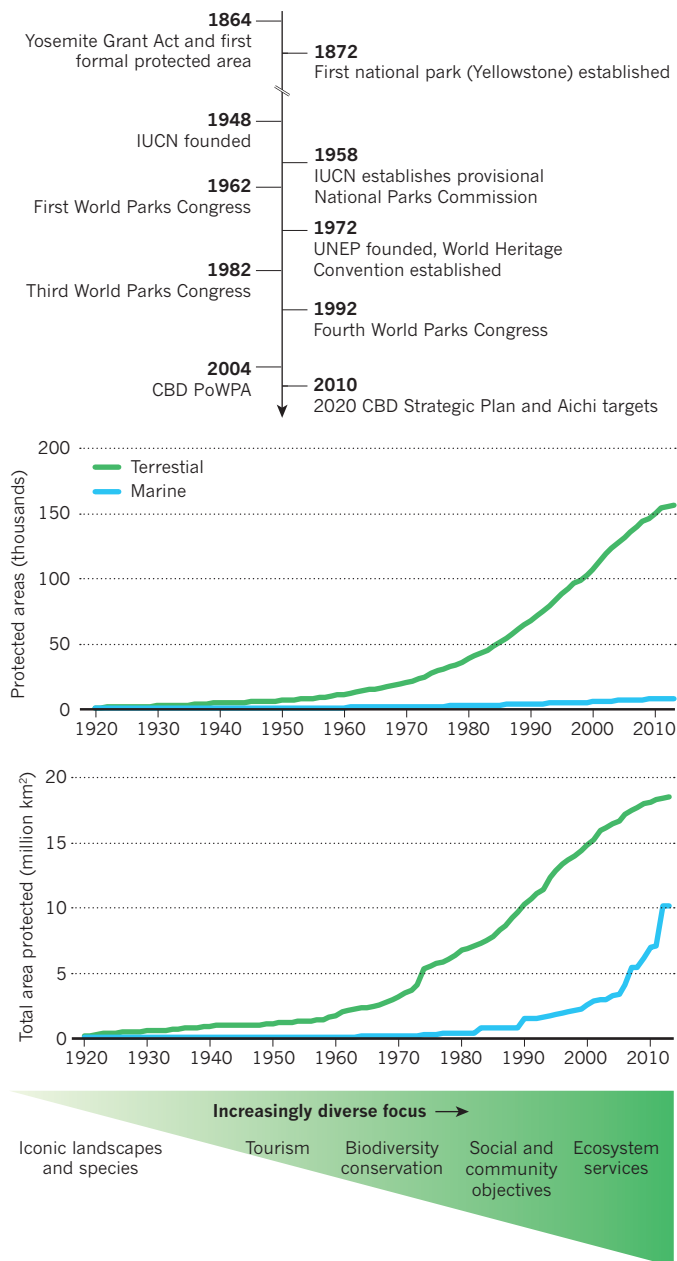
During the past two decades we have witnessed a greater emphasis on the role that functioning ecosystems have in maintaining human societies<sup>34</sup>. Well-managed protected areas can provide crucial ecosystem services, including water, food security, protection of wild relatives of crops, maintenance of wild fish stocks and carbon storage<sup>35–38</sup>. As populations urbanize, the role that protected areas have in providing clean water for cities is increasing: a third of the world's 100 largest cities rely on protected areas as a significant source of drinking water<sup>39</sup>. Protected areas are also now seen as a crucial component of global climate change mitigation efforts<sup>38</sup>, and are likely to play an increasingly important part in REDD+ (reducing emissions from deforestation and forest degradation, plus the conservation, sustainable management and enhancement of forest carbon stocks) schemes.

Although there is a strong global consensus within the conservation community that the principal role of protected areas is nature conservation<sup>40</sup>, in practice they are expected to make much wider ecological, social and economic contributions to human society. It is not yet clear what the overall impact of these increasing demands on protected areas will be or whether additional demands will arise in the future<sup>4</sup>.

### Coverage and targets for terrestrial protected areas

As of April 2014 (ref. 1), the official global portfolio of nationally designated terrestrial protected areas numbered 155,584 and covered 18.4 million km<sup>2</sup>, or 12.5% of the terrestrial realm (Supplementary Methods). This is still well short of the current CBD target of 17%<sup>6</sup>, a figure that has grown from 10% since the 2000–2010 strategic plan, but is still a political compromise that many conservation scientists believe is too low<sup>41</sup>. The shortfall is larger still when you consider CBD guidance that states protected areas should target places of “importance for biodiversity” that are “ecologically representative”<sup>6</sup>. At broad ecological scales, coverage markedly varies between major terrestrial biomes and ecoregions<sup>42</sup>. Using the latest available data<sup>1</sup>, we found that only 300 terrestrial ecoregions (36%) have more than 17% coverage, with 237 regions (29%) having less than 5% coverage and 68 (8%) having less than 1% coverage (Fig. 2a). When finer-scale analyses are conducted to assess whether protected areas are being placed in areas important for conserving species, the same patterns of variability occur. Among key biodiversity areas (KBAs)<sup>43</sup>, only 28% of Important Bird Areas (sites identified as crucial for bird biodiversity) and 22% of Alliance for Zero Extinction sites (sites that hold more than 95% of the global population of an endangered species<sup>44</sup>) are adequately covered by existing protected areas<sup>16</sup>. A recent global analysis of all threatened birds, amphibians and mammals ( $n = 4,118$ ) found that 17% are not found in a single protected area and 85% do not have sufficiently large populations in protected areas to give them a reasonable chance of long-term survival<sup>45</sup>. In comparison, a decade ago 20% of globally threatened terrestrial birds, mammals and amphibians were not found in a single protected area and 89% were inadequately represented<sup>46</sup>.

In the past, the patchy representation of species and ecosystems in



**Figure 1 | Growth of the modern terrestrial and marine protected area estate.** Growth in protected areas is aligned with a series of key events that have signalled an expansion of objectives over the past 150 years, starting with the establishment of the first formal protected area in 1864 (see Supplementary Table 1). New and increasingly diverse focal objectives have added to, rather than replaced, pre-existing objectives so that the requirements for management of protected areas have expanded over time. The growth was calculated using data obtained from the World Database on Protected Areas<sup>1</sup>.



protected areas has often been attributed to weaknesses in planning methods<sup>47</sup>. But these new analyses show that significant additions to the terrestrial protected area estate over the past two decades (Fig. 1) have not significantly lessened biases towards higher elevations, steeper slopes, and lands of lower productivity, lower economic worth and low human density<sup>48</sup>. Progress in achieving ecological representation has almost come to a stop, and this is likely to have serious ramifications when it comes to threats such as climate change<sup>49,50</sup>.

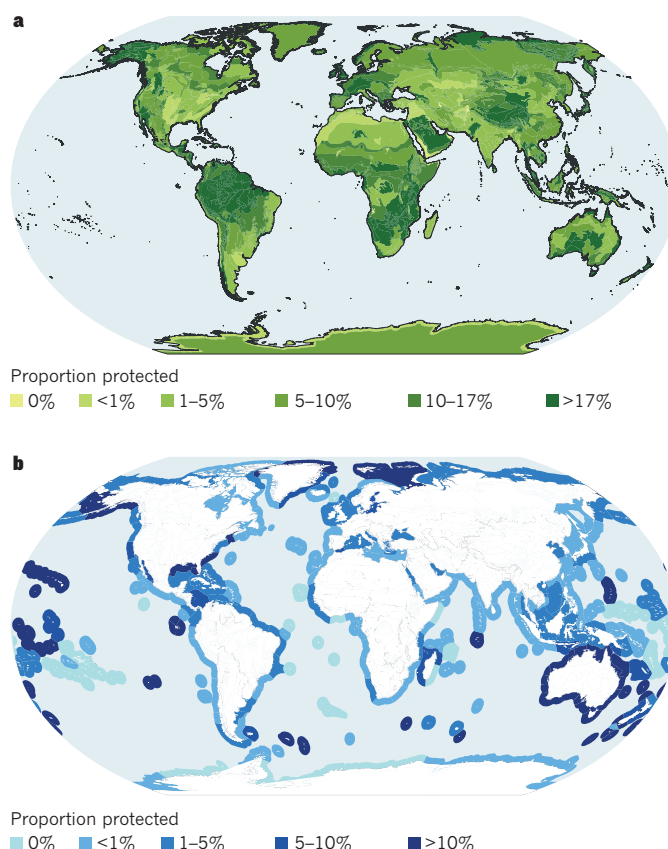
### Coverage and targets for marine protected areas

Progress towards the CBD's target of 10% of coastal and marine areas being protected<sup>6</sup> has been much slower than progress towards their terrestrial equivalent (Fig. 1). The goal is also much less ambitious than the 15–20% target proposed by non-governmental organizations (NGOs) and other groups attending the 2010 CBD Conference of the Parties. The tradition of open access to marine systems is a possible reason why it has taken longer for MPAs to be embraced than terrestrial protected areas. However, growth rates have been increasing since the World Summit on Sustainable Development in 2002 (ref. 51), and it is estimated that the 7,318 MPAs cover about 10.1 million km<sup>2</sup> (or 3%) of the world's marine environments<sup>1</sup>. Most designated MPAs have been restricted to national waters, with MPAs now covering about 6.6% of delimited exclusive economic zones globally<sup>52</sup>. The recent trend towards establishing very large (greater than 100,000 km<sup>2</sup>) MPAs is accelerating the expansion of global coverage: a recent analysis found that ten existing or under-creation MPAs account for more than 53% of the world's total MPA coverage<sup>53</sup>. Although large and remote MPAs may be important for maintaining functioning marine ecosystems, they will not avert imminent and direct anthropogenic threats in populated, coastal waters in which pressures on biodiversity often remain intense<sup>53</sup>. With these biases in mind, it is unsurprising that only 46 (20%) of the 232 marine ecoregions have more than 10% coverage and 107 ecoregions (46%) have less than 1% coverage (Fig. 2b). To our knowledge, there has been no fine-scale global analysis of MPA coverage of species or important marine biodiversity areas; this means there is a significant shortfall in our understanding of protected area gaps in the marine realm.

### Protected area effectiveness

Beyond achieving a percentage protected area coverage in the marine and terrestrial realm, the latest CBD targets<sup>6</sup> also call for protected areas to be effectively managed. Although systematic research into protected area effectiveness is still in its infancy, there are global studies that point to a significant shortfall in effectiveness — only 20–50% of protected areas assessed were found to be effectively managed<sup>26,54,55</sup>. These global averages mask even more critical situations in some ecosystems, and there are now many examples of protected areas not achieving basic objectives. For example, a recent assessment of coral reefs within MPAs in the western Pacific Ocean's Coral Triangle found that only 1% were effectively managed<sup>56</sup>, and research examining vegetation loss in protected areas in South Asia has shown that the trajectories of habitat conversion rates inside protected areas are sometimes indistinguishable from those of unprotected lands<sup>57</sup>. There is also evidence that threatened species populations inside some protected areas are declining<sup>58</sup>, including charismatic fauna such as lions (*Panthera leo*)<sup>59</sup>, Sumatran rhinoceros (*Dicerorhinus sumatrensis*)<sup>60</sup> and African elephants (*Loxodonta africana*)<sup>61</sup>.

Even some globally renowned protected areas, formally designated as UNESCO (United Nations Educational, Scientific and Cultural Organization) World Heritage Sites, have been shown to have experienced serious ecological degradation, partly due to poor management effectiveness. Australia's Great Barrier Reef Marine Park, for example, has experienced significant degradation, with large declines in coral cover in less than 30 years, as well as substantial declines in species populations, and habitat condition and extent across large areas of the park<sup>62,63</sup>. Similarly, Ecuador's Galapagos National Park and Marine Reserve — one of the best-known protected areas in the world — has widespread



**Figure 2 | Percentage of each terrestrial and marine ecoregion represented in the 2014 protected area estate.** **a**, The shortfall in protected area coverage of terrestrial ecoregions ( $n = 827$ ) relative to the Convention on Biological Diversity (CBD) target of 17%. As of 2014, only 300 ecoregions (36%) have more than 17% coverage, with 68 (8%) having less than 1% coverage and 237 (29%) of all ecoregions having less than 5% coverage. **b**, The shortfall in protected area coverage of marine ecoregions ( $n = 232$ ) relative to the CBD target of 10%. As of 2014, only 46 (20%) of the marine ecoregions have more than 10% coverage with 107 (46%) having less than 1%. (See Supplementary Methods for details of protected area calculations.)

problems including alien species invasion, population collapse of exploited marine species<sup>64</sup> and declines in ecosystem condition<sup>65</sup>.

Under-resourcing of protected area management is the primary reason for poor performance in protected area effectiveness, especially in the developing world<sup>66</sup>. Fewer than 6% of the countries reporting to the CBD in 2003 indicated that resources for management of protected areas were adequate, and it is unlikely this number has improved substantially<sup>67</sup>. A lack of resources affects boundary demarcation, effective law enforcement, natural and cultural resource management, and the provision of adequate park infrastructure, all of which affect protected area performance<sup>54,66</sup>. Effectiveness is further undermined by poor governance quality and bureaucratic inefficiency in many protected areas, alienating stakeholders and eroding support for management decisions<sup>68</sup>. Political corruption and armed conflict also undermines protected areas in many parts of the world, rendering protection efforts ineffective. Addressing these issues is likely to be a pre-condition for successfully instituting other reforms to improve management<sup>69</sup>.

Resources currently available for management pale in comparison with the challenges a protected area faces. Underlying pressures such as demographic growth, climate change and human consumption of natural resources, are all increasing, contributing to direct drivers of damage from encroachment, agriculture, infrastructure projects and timber demand<sup>70</sup>. These pressures are on the increase around many protected areas, with important implications for the overall effectiveness of the protected area — there are generally strong, positive correlations

between threats outside and inside protected areas<sup>71</sup>. Mining is now a serious pressure, with a recent global analysis showing that mining activity inside protected areas collectively affects 6% of the terrestrial protected area estate, and mines within 10 km of parks influence 14% of the global estate<sup>72</sup>. In the face of mounting threats, protected-area managers are hampered by the linked problems of insufficient resources, lack of capacity and, in many cases, a poor understanding about how to address the suite of pressures.

### Declining support for protected areas

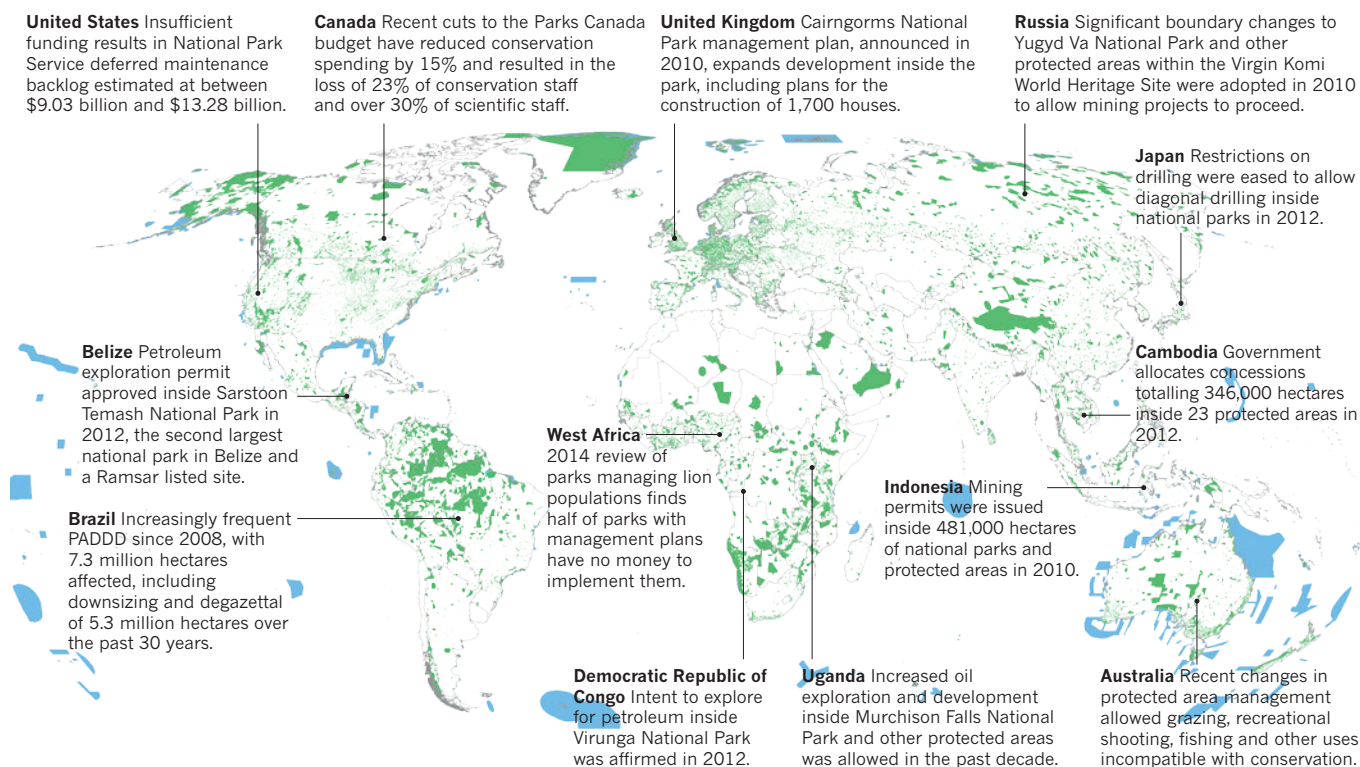
Despite making global commitments towards increasing the size and effectiveness of the protected area estate within the CBD framework<sup>6</sup>, there is now significant evidence that some governments are sliding back on their commitment to support protected areas through disproportionate funding cuts, reductions in professional staff and by ignoring their own policies (Fig. 3, Supplementary Table 2). If this is representative of a global trend, many protected areas will be left seriously exposed, especially in the context of pre-existing levels of underfunding and rising threats.

Although more common in developing countries<sup>73</sup>, inadequate financing of protected areas is also becoming increasingly prevalent in some of the richest countries, such as Australia, the United States and Canada, where major reductions in staffing levels and funding have been recently observed (Supplementary Table 2). For example, between 2009 and 2013, the US National Parks Service's main operational budget for national parks fell by nearly 13% (excluding the supplemental funding provided in response to Hurricane Sandy)<sup>74</sup>. Chronic underfunding has led to a deferred maintenance backlog estimated at between \$9.03 billion and \$13.28 billion, which has continued to build over the past decade<sup>74</sup>. In Canada, budget cuts have severely affected the science capacity of Parks Canada; ecosystem-science positions were cut by up to 30%, even though an expert panel, which was set up to review the

integrity of Canada's national parks, had identified the need for greater investment in this area to redress what they identified as significant impairments to the ecological condition of protected areas<sup>75</sup>.

Arguably, a more intransigent form of protected-area failure is government changes, through policy, that open up sites to resource extraction, or partial or full 'degazettement'. This practice has been labelled protected area downgrading, downsizing and degazettement (PADDD)<sup>76</sup>, where downgrading is the legal authorization of an increase in the number, magnitude or extent of human activities within a protected area; downsizing is the decrease in size of a protected area through a legal boundary change; and degazettement is the loss of legal protection for an entire protected area. A recent global analysis of 543 instances of PADDD indicated that all three forms are increasing<sup>77</sup>.

Most instances of PADDD have occurred in developing countries, where demographic changes and demand for land have put pressure on ecosystems, and where governments seek revenue from natural resources to meet development needs. It is not uncommon for the ministries responsible for mining or logging to issue leases on land or water that are already designated as protected. For example, in Uganda, active oil exploration and development is occurring inside protected areas, including the western portion of Murchison Falls National Park and inside three wildlife reserves that contain threatened species such as lions and Rothschild's giraffe (*Giraffa camelopardalis rothschildi*)<sup>78</sup>. Examples of downsizing include the World Heritage Sites the Selous Game Reserve in Tanzania, which was reduced to allow for uranium mining, and the Virgin Komi Forests in Russia where significant boundary changes to reserves such as the Yugyd Va National Park were made to allow mining (Supplementary Table 2). These are not isolated cases; a review of oil and gas concessions and protected areas in sub-Saharan Africa found that concessions overlapped park boundaries in 17.3% of International Union for Conservation of Nature (IUCN) category I–II sites and in more than a quarter of world heritage sites<sup>79</sup>.



**Figure 3 | A global portrait of different forms of decline in government support in terms of protected areas.** Here we provide examples of recent (defined as the past 10 years) decisions made by governments that are leading to a decline in the effectiveness of protected area estates within countries. The blue and green areas show the spatial extent of the global protected area

estate as of April 2014 (ref. 1). Examples are well-documented instances of this retreat; it should be noted that many more examples remain poorly documented. More information and additional examples are provided in Supplementary Table 2. PADDD, protected area downgrading, downsizing and degazettement.



Full degazettement is a more unusual form of PADD, but is also on the increase (Fig. 3; Supplementary Table 2): the Arabian Oryx Sanctuary in Oman was removed from the World Heritage List after the government reduced the size of the reserve by 90% to allow for oil and gas extraction<sup>80</sup>. Developed countries are not immune to PADD, as witnessed in Australia in recent years with the opening up of parks to allow industrial logging, livestock grazing, mining, recreational hunting and fishing<sup>81</sup>. In the United Kingdom, a proposed high-speed rail link will damage or destroy ten Sites of Special Scientific Interest, government-recognized protected areas and nine NGO-run Wildlife Trust reserves<sup>82</sup>.

### The case for a step change

The current biodiversity crisis would be much more severe had the establishment of protected areas in the last century not occurred. But a fundamental increase in support of the global protected area estate is now urgently needed if it is to fully deliver its potential. Countries are still well short of what they formally agreed to do in the 2020 CBD strategic plan<sup>6</sup>, and the resources currently provided for protected-area management are, for the most part, insufficient. Many countries are even failing to take adequate steps to maintain what they have already designated for protection. The increasing number of governments overtly or covertly decreasing resources and allowing incompatible use of protected areas sounds a clear alarm bell for current and future performance.

This erosion of support for maintaining and growing the global protected area estate is occurring at a time when anthropogenic climate change, the sudden upsurge in poaching, 'land-grabbing' by powerful businesses and increased mining activity are all making conservation challenges more complex. As human populations grow and pressures on natural ecosystems increase, more species and ecosystems are becoming predominantly (and in some cases, completely) confined to protected areas. These include large, threatened mammals such as the Asian elephant (*Elephas maximus*)<sup>83</sup>, the tiger (*Panthera tigris*)<sup>27</sup> and all rhinoceros species<sup>84</sup>, but also numerous plants, reptiles and amphibians<sup>44,85</sup>. At the same time, large-scale anthropogenic modification of natural ecosystems means that protected areas are now crucial for sustaining a large proportion of the world's poorest people by providing them with life's most basic subsistence necessities — food, water, shelter and medicine<sup>3</sup>.

A business-as-usual approach, by which most countries do not provide adequate resources to ensure effective management of protected areas and undervalue the need for continued expansion of the protected area estate, means that the broad goals of the estate will fail. A fundamental step change is needed to ensure both the current and future potential of the estate is met.

First, countries need to create management regimes for existing protected areas that ensure they are effective together with policies that support protected-area systems. Management outside protected areas can, and must, make crucial contributions to securing the future of biodiversity, but this is in no way a replacement for the clear benefits that protected areas provide, especially with respect to conserving KBAs<sup>43</sup>. Crucially, policies should be aligned within government so that the actions of ministries dealing with development, resource extraction and agriculture do not undermine those of ministries concerned with the environment and conservation. Much progress could be achieved simply by implementing approved policies on protected areas that have already been established through multilateral environmental agreements such as the CBD, the Ramsar Convention on Wetlands and the World Heritage Convention, as well as implementing the IUCN's guidance on the definition, management objectives and governance types within protected areas<sup>40</sup>. Adopting the legal principle of non-regression when it comes to environmental laws is also an important action that countries can take<sup>86</sup>. Another part of the step change is for countries to acknowledge that the widening expectations of protected area systems, along with growing pressures on protected areas, increase the burden for all those involved in management and requires improvements rather than declines in policy and resource support.

Second, countries need to invest adequately in protected areas to ensure that their objectives are achieved. Part of this is recognizing the return on investment that well-managed protected areas provide by conserving natural heritage and increasing the social and economic well-being of their citizens<sup>4,87</sup>. Countries need to start quantifying the services provided by protected areas and recognizing the costs of protected-area degradation — it should be noted that many of the benefits will not easily be measurable in monetary terms. In Australia, the 2012–13 budget for the Great Barrier Reef Marine Park Authority was approximately Aus\$50 million, but tourism to the reef was worth more than Aus\$5.2 billion annually to the Australian economy<sup>62</sup>; this income is seriously threatened by the current degradation of the reef. In 2009, the Canadian government spent Can\$800 million, but the contribution to the economy was Can\$4.6 billion and supported the employment of 64,000 people<sup>75</sup>. A better understanding of the returns on investment would help to persuade countries of the need to provide resources to protected areas that better match the benefits received. In the Eastern Arc Mountains in East Africa, one-third of protected areas do not receive the minimum funds necessary to be effective reserves, even though it would only require reinvestment of 13% of the revenue raised through protected area tourism in Tanzania to fully fund the protected area estate<sup>88</sup>.

There is ample evidence to justify more state support of protected areas. But good arguments do not always translate into large amounts of financial resources, and conservation is often an early casualty of any government funding squeeze. As a consequence, the third component of the step change is to accept the fact that governments will often not supply sufficient financial resources for protected areas and that there is a need to identify innovative models for ensuring protected area success; in other words, to encourage the wider community to take collective responsibility for protected areas. Non-conventional funding sources (philanthropic contributions, and payments for ecosystem service mechanisms such as REDD+) have the potential to be crucially important future alternative funding sources, as do mechanisms such as offsets and 'debt-for-nature swaps' from the corporate sector<sup>89</sup>. All these options need careful appraisal and, where appropriate, more strategic application; we know that some of them can work in some places, but they are currently applied very sporadically<sup>90,91</sup>. In addition to broadening the funding base of protected areas, the next required change of approach is for a similar explosion in management collaborations. Building resilient social constituencies that advocate on behalf of protected areas and biodiversity conservation requires the formation of coalitions across local, national and international actors (government, NGO, business and community groups) across the political spectrum; yet too little attention has been paid to the requirements of building a political constituency that will underpin long-lasting commitments to conservation<sup>92</sup>. At the site scale, volunteers can provide a substantial resource for managers of protected areas, filling roles ranging from those involving site maintenance to anti-poacher patrols, and there is evidence to show that properly planned voluntary efforts can be harmonized with professional activities to produce the desired management outcomes<sup>93</sup>. For example, partnerships between protected area agencies and scientists can bridge research and monitoring gaps in a mutually beneficial way, and citizen science can not only provide information for managers but also build a supportive and hopeful constituency for conservation<sup>94–96</sup>.

Finally, most countries still need to expand their protected area networks to meet CBD obligations and the pressing needs outlined in this Review. The challenges of achieving this on an increasingly crowded planet should not be underestimated. It will require countries to embrace transparent planning frameworks to identify the new areas needed to achieve the objectives outlined clearly by the CBD, rigorous stakeholder consultation, imaginative application of a range of management approaches and then acting on plans and monitoring results transparently. A key element is to expand reporting beyond simply the area of land and seas gazetted to include ecological connectivity, management

effectiveness, equity, social and economic benefits, and the contribution of the system to conserving areas that are important for biodiversity.

The package of responses needed for the step change is neither impossible nor unreasonable, although individual countries may struggle with some of the components. Fundamentally, it requires the recognition that protected areas are core to the future of life on our planet. Estimations of the annual cost of adequately managing an expanded network of marine and terrestrial protected areas range from \$45 billion to \$76 billion<sup>97,98</sup>, the lower of which is just 2.5% of the global military expenditure<sup>99</sup>. But adequate protection of marine and terrestrial environments is also crucial to global security. It seems sensible to invest an amount equivalent to a tiny percentage of global military spending to help provide security for humans and all other living organisms on Earth through a system of marine and terrestrial protected areas that is operating at its full capacity. Although we need to understand and report on the performance of protected areas, we also need to focus on the promise and potential they provide for the well-being of the planet and its inhabitants. ■

Received 6 March; accepted 26 June 2014.

1. UNEP-WCMC. *World Database on Protected Areas* <http://www.wdpa.org> (accessed April 2014).
2. Sanderson, E. W. *et al.* The human footprint and the last of the wild. *Bioscience* **52**, 891–904 (2002).
3. Stolton, S. & Dudley, N. *Arguments for Protected Areas: Multiple Benefits for Conservation and Us* (Earthscan, 2010).
4. Naughton-Treves, L., Holland, M. B. & Brandon, K. The role of protected areas in conserving biodiversity and sustaining local livelihoods. *Annu. Rev. Environ. Resour.* **30**, 219–252 (2005).
5. **This is an important review that explores the different roles of protected areas and the implications of different objectives.**
6. White, C., Halpern, B. S. & Kappel, C. V. Ecosystem service tradeoff analysis reveals the value of marine spatial planning for multiple ocean uses. *Proc. Natl Acad. Sci. USA* **109**, 4696–4701 (2012).
7. Convention on Biological Diversity. *COP 10 Decision X/2: Strategic Plan for Biodiversity 2011–2020* <http://www.cbd.int/decision/cop/?id=12268> (CBD, 2011).
8. **The CBD strategic plan for protected areas identifies targets that countries are obligated to meet by 2020.**
9. Chape, S., Harrison, J., Spalding, M. & Lysenko, I. Measuring the extent and effectiveness of protected areas as an indicator for meeting global biodiversity targets. *Phil. Trans. R. Soc. B* **360**, 443–455 (2005).
10. Shahabuddin, G. & Rao, M. Do community-conserved areas effectively conserve biological diversity? Global insights and the Indian context. *Biol. Conserv.* **143**, 2926–2936 (2010).
11. **This is the first large review of community conserved areas and how they compare with other forms of protected areas.**
12. Peres, C. A. & Nascimento, H. S. Impact of game hunting by the Kayapó of south-eastern Amazonia: implications for wildlife conservation in tropical forest indigenous reserves. *Biodivers. Conserv.* **15**, 2627–2653 (2006).
13. Phillips, A. The history of the international system of protected area management categories. *Parks* **14**, 4–14 (2004).
14. **This paper provides an important history of the modern protected area movement.**
15. Runte, A. The national park idea: origins and paradox of the American experience. *J. For. Hist.* **21**, 64–75 (1977).
16. Zeiger, J. B., Caneday, L. M. & Baker, P. R. Symbiosis between tourism and our national parks. *Parks Recreat.* **27**, 74–79 (1992).
17. Balmford, A. *et al.* A global perspective on trends in nature-based tourism. *PLoS Biol.* **7**, e1000144 (2009).
18. Maekawa, M., Lanjouw, A., Rutagarama, E. & Sharp, D. Mountain gorilla tourism generating wealth and peace in post-conflict Rwanda. *Nat. Resour. Forum* **37**, 127–137 (2013).
19. International Union for Conservation of Nature. *World Conservation Strategy: Living Resource Conservation for Sustainable Development* (IUCN–UNEP–WWF, 1980).
20. Butchart, S. H. M. *et al.* Protecting important sites for biodiversity contributes to meeting global conservation targets. *PLoS ONE* **7**, e32529 (2012).
21. Geldmann, J. *et al.* Effectiveness of terrestrial protected areas in reducing habitat loss and population declines. *Biol. Conserv.* **161**, 230–238 (2013).
22. Joppa, L. & Pfaff, A. Reassessing the forest impacts of protection. *Ann. NY Acad. Sci.* **1185**, 135–149 (2010).
23. Micheli, F. & Niccolini, F. Achieving success under pressure in the conservation of intensely used coastal areas. *Ecol. Soc.* **18**, 19 (2013).
24. Edgar, G. J. *et al.* Global conservation outcomes depend on marine protected areas with five key features. *Nature* **506**, 216–220 (2014).
25. Lester, S. E. *et al.* Biological effects within no-take marine reserves: a global synthesis. *Mar. Ecol. Prog. Ser.* **384**, 33–46 (2009).
26. **This is an important global synthesis on the effectiveness of strict MPAs.**
27. Karanth, K. K., Nichols, J. D., Hines, J. E., Karanth, K. U. & Christensen, N. L. Patterns and determinants of mammal species occurrence in India. *J. Appl. Ecol.* **46**, 1189–1200 (2009).
28. Taylor, M. *et al.* What works for threatened species recovery? An empirical evaluation for Australia. *Biodivers. Conserv.* **20**, 767–777 (2011).
29. Sheehan, E. V., Stevens, T. F., Gall, S. C., Cousens, S. L. & Attrill, M. J. Recovery of a temperate reef assemblage in a marine protected area following the exclusion of towed demersal fishing. *PLoS ONE* **8**, e83883 (2013).
30. Sciberras, M., Jenkins, S. R., Kaiser, M. J., Hawkins, S. J. & Pullin, A. S. Evaluating the biological effectiveness of fully and partially protected marine areas. *Environ. Evid.* **2**, 4 (2013).
31. Laurance, W. F. *et al.* Averting biodiversity collapse in tropical forest protected areas. *Nature* **489**, 290–294 (2012).
32. **This article is an important meta-analysis of 60 protected areas located across the tropics that assess the different drivers of change in protected areas and differential impacts on species.**
33. Walston, J. *et al.* Bringing the tiger back from the brink — the six percent solution. *PLoS Biol.* **8**, e1000485 (2010).
34. Hilborn, R. *et al.* Effective enforcement in a conservation area. *Science* **314**, 1266 (2006).
35. Brockington, D., Igoe, J. & Schmidt-Soltau, K. Conservation, human rights and poverty reduction. *Conserv. Biol.* **20**, 250–252 (2006).
36. Agrawal, A. & Redford, K. Conservation and displacement: an overview. *Conserv. Soc.* **7**, 1 (2009).
37. Dudley, N. *et al.* in *Partnerships for Protection: New Strategies for Planning Management for Protected Areas* (Stolton, S. & Dudley, N.) 3–12 (Earthscan, 1999).
38. Ferraro, P. J. & Hanauer, M. M. Protecting ecosystems and alleviating poverty with parks and reserves: ‘win-win’ or tradeoffs? *Environ. Resour. Econ.* **48**, 269–286 (2011).
39. Ferraro, P. J., Hanauer, M. M. & Sims, K. R. E. Conditions associated with protected area success in conservation and poverty reduction. *Proc. Natl Acad. Sci. USA* **108**, 13913–13918 (2011).
40. Costanza, R. *et al.* The value of the world’s ecosystem services and natural capital. *Nature* **387**, 253–260 (1997).
41. Postel, S. L. & Thompson, B. H. Watershed protection: capturing the benefits of nature’s water supply services. *Nat. Resour. Forum* **29**, 98–108 (2005).
42. Lubchenco, J., Palumbi, S. R., Gaines, S. D. & Andelman, S. Plugging a hole in the ocean: the emerging science of marine reserves. *Ecol. Appl.* **13**, 3–7 (2003).
43. Scharlemann, J. P. *et al.* Securing tropical forest carbon: the contribution of protected areas to REDD. *Oryx* **44**, 352–357 (2010).
44. Soares-Filho, B. *et al.* Role of Brazilian Amazon protected areas in climate change mitigation. *Proc. Natl Acad. Sci. USA* **107**, 10821–10826 (2010).
45. Dudley, N. & Stolton, S. *Running Pure: the Importance of Forest Protected Areas to Drinking Water* (World Bank/WWF Alliance for Forest Conservation and Sustainable Use, 2003).
46. International Union for Conservation of Nature. *Guidelines for Applying Protected Area Management Categories* (IUCN, 2008).
47. Noss, R. F. *et al.* Bolder thinking for conservation. *Conserv. Biol.* **26**, 1–4 (2012).
48. Jenkins, C. N. & Joppa, L. Expansion of the global terrestrial protected area system. *Biol. Conserv.* **142**, 2166–2174 (2009).
49. Eken, G. *et al.* Key biodiversity areas as site conservation targets. *Bioscience* **54**, 1110–1118 (2004).
50. Ricketts, T. H. *et al.* Pinpointing and preventing imminent extinctions. *Proc. Natl Acad. Sci. USA* **102**, 18497–18501 (2005).
51. **This was the first global assessment identifying sites that hold crucial populations of an endangered or critically endangered species.**
52. Venter, O. *et al.* Targeting global protected area expansion for imperiled biodiversity. *PLoS Biol.* **12**, e1001891 (2014).
53. **This article provides an important analysis showing the shortfall of the global protected area estate in protecting threatened species with key recommendations for the CBD.**
54. Rodrigues, A. S. L. *et al.* Effectiveness of the global protected area network in representing species diversity. *Nature* **428**, 640–643 (2004).
55. Pressey, R. L. Ad hoc reservations: forward or backward steps in developing representative reserve systems. *Conserv. Biol.* **8**, 662–668 (1994).
56. Watson, J. E. M. *et al.* Wilderness and future conservation priorities in Australia. *Divers. Distrib.* **15**, 1028–1036 (2009).
57. Hannah, L. *et al.* Protected area needs in a changing climate. *Front. Ecol. Environ.* **5**, 131–138 (2007).
58. Watson, J. E. M., Iwamura, T. & Butt, N. Mapping vulnerability and conservation adaptation strategies under climate change. *Nature Clim. Change* **3**, 989–994 (2013).
59. Wood, L. J., Fish, L., Laughren, J. & Pauly, D. Assessing progress towards global marine protection targets: shortfalls in information and action. *Oryx* **42**, 340–351 (2008).
60. Ardrón, J., Gjerde, K., Pullen, S. & Tilot, V. Marine spatial planning in the high seas. *Mar. Policy* **32**, 832–839 (2008).
61. Devillers, R. *et al.* Reinventing residual reserves in the sea: are we favouring ease of establishment over need for protection? *Aquat. Conserv. Mar. Freshw. Ecosyst.* <http://dx.doi.org/10.1002/aqc.2445> (2014).
62. Leverington, F., Costa, K. L., Pavese, H., Lisle, A. & Hockings, M. A global analysis of protected area management effectiveness. *Environ. Manage.* **46**, 685–698 (2010).
63. **This paper reports the first attempt at a global assessment of protected area effectiveness.**
64. Blom, A., Yamindou, J. & Prins, H. H. T. Status of the protected areas of the Central African Republic. *Biol. Conserv.* **118**, 479–487 (2004).
65. Burke, L. M., Reyter, K., Spalding, M. & Perry, A. *Reefs at Risk Revisited in the Coral Triangle* (WRI, 2012).



57. Clark, N. E., Boakes, E. H., McGowan, P. J. K., Mace, G. M. & Fuller, R. A. Protected areas in South Asia have not prevented habitat loss: a study using historical models of land-use change. *PLoS ONE* **8**, e65298 (2013).
58. Craigie, I. D. *et al.* Large mammal population declines in Africa's protected areas. *Biol. Conserv.* **143**, 2221–2228 (2010).
59. Riggio, J. *et al.* The size of savannah Africa: a lion's (*Panthera leo*) view. *Biodivers. Conserv.* **22**, 17–35 (2013).
60. Ahmad Zafir, A. W. *et al.* Now or never: what will it take to save the Sumatran rhinoceros *Dicerorhinus sumatrensis* from extinction? *Oryx* **45**, 225–233 (2011).
61. Maisels, F. *et al.* Devastating decline of forest elephants in Central Africa. *PLoS ONE* **8**, e59469 (2013).
62. Great Barrier Reef Marine Park Authority. *Great Barrier Reef Region Strategic Assessment: Strategic Assessment Report* <http://hdl.handle.net/11017/2861> (Great Barrier Reef Marine Park Authority, 2014).
63. Brodie, J. & Waterhouse, J. A critical review of environmental management of the 'not so Great' Barrier Reef. *Estuar. Coast. Shelf Sci.* **104–105**, 1–22 (2012).
64. Bucaram, S. J., White, J. W., Sanchirico, J. N. & Wilen, J. E. Behavior of the Galapagos fishing fleet and its consequences for the design of spatial management alternatives for the red spiny lobster fishery. *Ocean Coast. Manage.* **78**, 88–100 (2013).
65. Watson, J., Trueman, M., Tufet, M., Henderson, S. & Atkinson, R. Mapping terrestrial anthropogenic degradation on the inhabited islands of the Galapagos Archipelago. *Oryx* **44**, 79–82 (2010).
66. Bruner, A. G., Gullison, R. E., Rice, R. E. & da Fonseca, G. A. B. Effectiveness of parks in protecting tropical biodiversity. *Science* **291**, 125–128 (2001).  
**This was one of the first large-scale analyses on what makes protected areas effective in protecting biodiversity.**
67. Convention on Biodiversity. *Protected areas: Synthesis of Information in Thematic Reports on Protected Areas: Note by the Executive Secretary* (CBD, 2003).
68. Borri-Feyerabend, G. *et al.* *Governance of Protected Areas: From Understanding to Action* (IUCN, 2013).
69. Irland, L. C. State failure, corruption, and warfare: challenges for forest policy. *J. Sustain. For.* **27**, 189–223 (2008).
70. Spalding, M. & McManus, E. in *The Worlds Protected Areas: Status, Values and Prospects in the 21st Century* (eds Chape, S., Spalding, M. & Jenkins, M.) 146–157 (Univ. California Press, 2008).
71. DeFries, R., Hansen, A., Newton, A. C. & Hansen, M. C. Increasing isolation of protected areas in tropical forests over the past twenty years. *Ecol. Appl.* **15**, 19–26 (2005).
72. Durán, A. P., Rauch, J. & Gaston, K. J. Global spatial coincidence between protected areas and metal mining activities. *Biol. Conserv.* **160**, 272–278 (2013).
73. Bovarnick, A., Fernandez-Baca, J., Galindo, J. & Negret, H. *Financial Sustainability of Protected Areas in Latin America and the Caribbean: Investment Policy Guidance* (The Nature Conservancy and UNDP, 2010).
74. Comay, L. B. *National Park Service: Recent Appropriations Trends* (Congressional Research Service, 2013).
75. Canadian Parks and Wilderness Society. *The State of Canada's Parks 2012* (Canadian Parks and Wilderness Society, 2012).
76. Mascia, M. B. & Pailler, S. Protected area downgrading, downsizing, and degazettement (PADDD) and its conservation implications. *Conserv. Lett.* **4**, 9–20 (2011).  
**This was the first global assessment documenting instances of PADDD.**
77. Mascia, M. B. *et al.* Protected area downgrading, downsizing, and degazettement (PADDD) in Africa, Asia, and Latin America and the Caribbean, 1900–2010. *Biol. Conserv.* **169**, 355–361 (2014).
78. Republic of Uganda. *National Oil And Gas Policy For Uganda* (Ministry of Energy and Mineral Development, 2008).
79. Osti, M., Coad, L., Fisher, J. B., Bomhard, B. & Hutton, J. M. Oil and gas development in the World Heritage and wider protected area network in sub-Saharan Africa. *Biodivers. Conserv.* **20**, 1863–1877 (2011).
80. Frey, B. S. & Steiner, L. World Heritage List: does it make sense? *Int. J. Cult. Policy* **17**, 555–573 (2011).
81. Ritchie, E. G. *et al.* Continental-scale governance and the hastening of loss of Australia's biodiversity. *Conserv. Biol.* **27**, 1133–1135 (2013).
82. Watkins, H., Hawkins, K. & Cormack, A. *HS2: The Case for a Greener Vision* (The Wildlife Trusts, 2014).
83. Sukumar, R. A brief review of the status, distribution and biology of wild Asian elephants *Elephas maximus*. *Int. Zoo Yearb.* **40**, 1–8 (2006).
84. Emslie, R. & Brooks, M. *African Rhino Status Survey and Conservation Action Plan* (ICUN/SSC African Rhino Specialist Group, 1999).
85. Le Saout, S. *et al.* Protected areas and effective biodiversity conservation. *Science* **342**, 803–805 (2013).
86. Prieur, M. *Non-regression in Environmental Law* <http://sapiens.revues.org/1405> (SAPIENS, 2012).
87. Willemsen, L., Drakou, E. G., Dunbar, M. B., Mayaux, P. & Egoh, B. N. Safeguarding ecosystem services and livelihoods: understanding the impact of conservation strategies on benefit flows to society. *Ecosyst. Serv.* **4**, 95–103 (2013).
88. Green, J. M. H. *et al.* Estimating management costs of protected areas: a novel approach from the Eastern Arc Mountains, Tanzania. *Biol. Conserv.* **150**, 5–14 (2012).
89. Edwards, D. P. *et al.* Mining and the African environment. *Conserv. Lett.* **7**, 302–311 (2014).
90. Knicley, J. E. Debt, nature, and indigenous rights: twenty-five years of debt-for-nature evolution. *HELR Harvard Environ. Law Rev.* **36**, 79–122 (2012).
91. Gross-Camp, N. D., Martin, A., McGuire, S., Kebede, B. & Munyarukaza, J. Payments for ecosystem services in an African protected area: exploring issues of legitimacy, fairness, equity and effectiveness. *Oryx* **46**, 24–33 (2012).
92. Steinberg, P. F. Institutional resilience amid political change: the case of biodiversity conservation. *Glob. Environ. Polit.* **9**, 61–81 (2009).
93. Tulloch, A. I. T., Mustin, K., Possingham, H. P., Szabo, J. K. & Wilson, K. A. To boldly go where no volunteer has gone before: predicting volunteer activity to prioritize surveys at the landscape scale. *Divers. Distrib.* **19**, 465–480 (2013).
94. Swaisgood, R. R. & Sheppard, J. K. The culture of conservation biologists: show me the hope! *Bioscience* **60**, 626–630 (2010).
95. Banerjee, K., Jhala, Y. V., Chauhan, K. S. & Dave, C. V. Living with lions: the economics of coexistence in the Gir Forests, India. *PLoS ONE* **8**, e89708 (2013).
96. Rastogi, A., Hickey, G. M., Badola, R. & Hussain, S. A. Saving the superstar: a review of the social factors affecting tiger conservation in India. *J. Environ. Manage.* **113**, 328–340 (2012).
97. McCarthy, D. P. *et al.* Financial costs of meeting global biodiversity conservation targets: current spending and unmet needs. *Science* **338**, 946–949 (2012).
98. Balmford, A. *et al.* Economic reasons for conserving wild nature. *Science* **297**, 950–953 (2002).
99. Stockholm International Peace Research Institute. *Military Expenditure* <http://www.sipri.org/research/armaments/milex> (SPIRI, 2014).

**Supplementary Information** is linked to the online version of the paper at [www.nature.com/nature](http://www.nature.com/nature).

**Acknowledgements** We thank N. Butt, M. Callow, T. Evans, M. Giampieri, J. Hilty, K. MacKinnon, T. McClanahan, M. Rao, E. Sanderson, T. Stevens, S. Stolton, K. Redford, J. Robinson, J. Walston and S. Woodley for their thoughtful comments on earlier versions of this manuscript. We thank B. MacSharry for supplying the latest WDPa protected area data and many colleagues within the IUCN WCPa who have provided information and advice. Because of the reference limitations for this Review, we recognize that many important references have not been cited.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of this paper at [go.nature.com/Ini9wc](http://go.nature.com/Ini9wc). Correspondence should be addressed to J.E.M.W. ([jwatson@wcs.org](mailto:jwatson@wcs.org)).

# Life cycles, fitness decoupling and the evolution of multicellularity

Katrin Hammerschmidt<sup>1\*</sup>, Caroline J. Rose<sup>1\*</sup>, Benjamin Kerr<sup>2</sup> & Paul B. Rainey<sup>1,3</sup>

**Cooperation is central to the emergence of multicellular life; however, the means by which the earliest collectives (groups of cells) maintained integrity in the face of destructive cheating types is unclear. One idea posits cheats as a primitive germ line in a life cycle that facilitates collective reproduction. Here we describe an experiment in which simple cooperating lineages of bacteria were propagated under a selective regime that rewarded collective-level persistence. Collectives reproduced via life cycles that either embraced, or purged, cheating types. When embraced, the life cycle alternated between phenotypic states. Selection fostered inception of a developmental switch that underpinned the emergence of collectives whose fitness, during the course of evolution, became decoupled from the fitness of constituent cells. Such development and decoupling did not occur when groups reproduced via a cheat-purging regime. Our findings capture key events in the evolution of Darwinian individuality during the transition from single cells to multicellularity.**

Cooperation has a central role in the evolution of multicellularity<sup>1–8</sup>. Even under laboratory conditions, simple undifferentiated groups of cooperating cells readily evolve; however, such groups are often short lived: selection typically favours the evolution of cheats<sup>9–14</sup>. Cheats are cells that do not contribute towards group integrity, but nonetheless take advantage of the benefit that accrues from being part of a collective<sup>7</sup>. In the absence of cheater-suppression mechanisms, cheats may prosper to the point where the integrity of any newly emerged group is compromised<sup>15–18</sup>. The problem of cheating has led to the suggestion that the evolution of mechanisms for cheater suppression is a critical step in the transition to multicellularity<sup>18–21</sup>. In this Article we explore an alternative possibility; namely, that cheats may play a critical role in a simple multicellular life cycle where the central problem is not cheater suppression, but rather controlled generation of this phenotype<sup>22</sup>.

The evolution of simple groups of bacteria occurs repeatedly when populations of the bacterium *Pseudomonas fluorescens* are propagated in spatially structured microcosms<sup>9,23,24</sup>. Such collectives—‘wrinkly spreader’ (WS) mats—arise by spontaneous mutations from the ancestral ‘smooth’ (SM) genotype<sup>23,25</sup>. The mutations cause WS cells to over-produce a cell–cell glue<sup>25–27</sup> that holds daughter cells together following cell division<sup>28</sup>. The net effect is a cellular mat that colonizes the air–broth interface. Although glue production is costly to individual cells, the trait spreads<sup>29</sup> because the group of mat-forming cells reaps an advantage (access to oxygen) that is denied to individual cells<sup>9</sup>.

The life span of WS mats is brief: selection acting on individual cells favours mutant types that cheat. Cheating cells are phenotypically SM and no longer produce adhesive glues<sup>9</sup>; nonetheless they take advantage of the benefit that accrues from being part of the mat. In the absence of any mechanism of cheater repression, cheats prosper—ultimately weakening the fabric of the mat to the point where it collapses<sup>9</sup>.

While cheats pose a significant problem, the tension between cooperating and cheating cells could fuel evolution<sup>22,30,31</sup>. Consider a newly emergent WS mat. In the absence of a means of collective-level reproduction, the mat, like soma, is an evolutionary dead end. The emergence of cheats however is guaranteed. While cheats may destroy the mat, they also stand as a means of mat reproduction, provided they can regenerate

the mat. The cycling between WS groups and SM cells—driven by the niche-constructing activities of each type<sup>32</sup>—has the potential to generate a primitive life cycle<sup>30</sup> (Fig. 1a, left panel) and with this the possibility that WS groups might participate, even just marginally, as units of selection in the process of Darwinian evolution<sup>33</sup>.

## Embracing cheats

To test the plausibility of the idea that SM cells might function as the seeds for a new generation of WS mats, we propagated mats under a regime in which SM cells were integral to mat reproduction (Fig. 1a, left panel) (this treatment stands in stark contrast to a regime discussed below in which cheats are purged (Fig. 1a, right panel)). Each generation was founded by a single WS genotype (Fig. 1b). For lines to persist WS mats had to remain intact (viable) until the end of phase I, and be fecund (produce SM types). In addition, during the three-day phase II period, SM cells were required to transition back to WS. On completion of the cycle a single colony of the most dominant WS type was transferred to a fresh microcosm ensuring new WS mats arise through a bottleneck and are thus re-established free of within-mat conflicts<sup>5,34</sup>.

Transitioning between group and single-cell phases relies initially upon mutation and poses significant challenges. In an initial experiment in which 120 lines were required to repeatedly transition between WS and SM states, all lines went extinct by the sixth life-cycle generation. Extinction was primarily due to insufficient SM-cell production before day six of phase I (Fig. 1c).

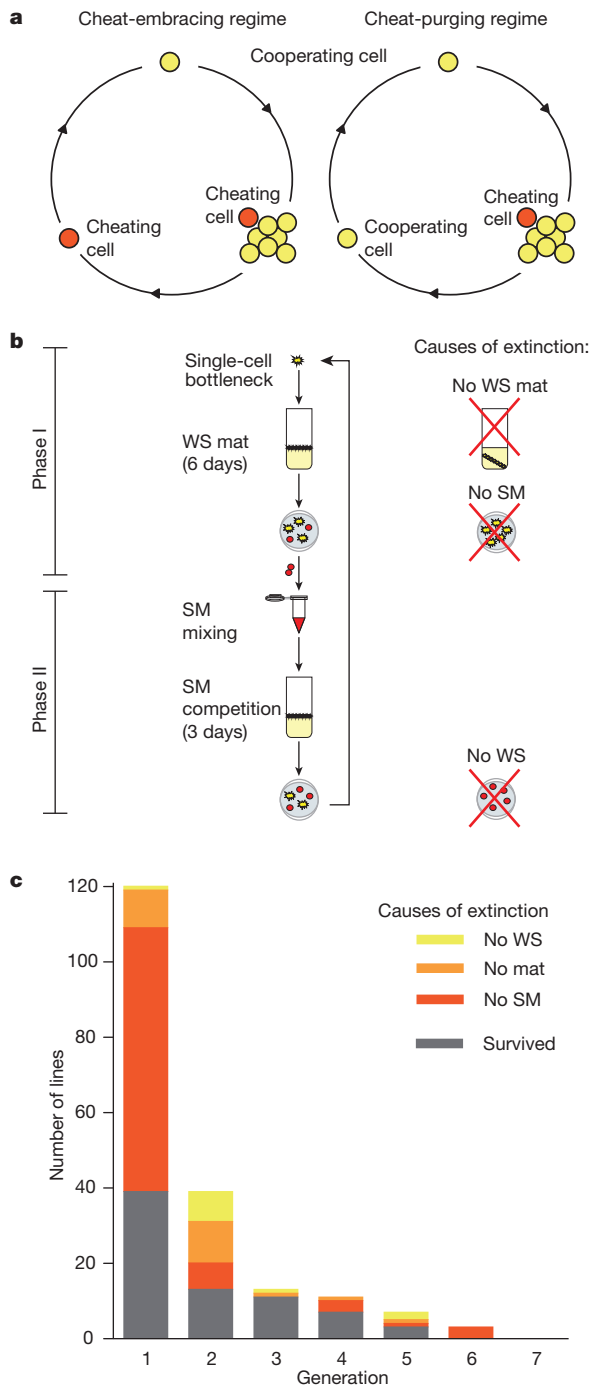
Given the prominent role of cheats and the requirement for mutation to transition each phase of the life cycle, the persistence of lines through six generations is surprising; it indicates a capacity for innovation. Such a capacity might, under different circumstances, provide opportunity for evolutionary refinement to the point where cycling through phases could come under developmental control.

Most non-neutral mutations are deleterious and thus prone to eventually disrupt the life cycle. Persistence is therefore likely to depend on viable lines having an opportunity to export their success to a new microcosm by division of the lineage. Were the splitting of viable lines to operate concomitantly with the elimination of unsuccessful lines then selection

<sup>1</sup>New Zealand Institute for Advanced Study and Allan Wilson Centre for Molecular Ecology & Evolution, Massey University, Auckland 0745, New Zealand. <sup>2</sup>Department of Biology and BEACON Center for the Study of Evolution in Action, University of Washington, Seattle, Washington 98195, USA. <sup>3</sup>Max Planck Institute for Evolutionary Biology, Plön 24306, Germany.

\*These authors contributed equally to this work.



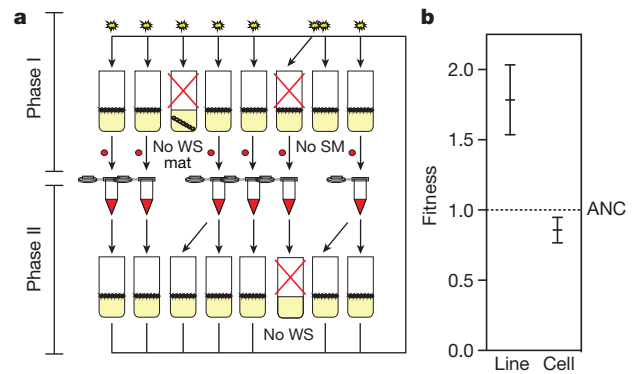


**Figure 1 | Experimental regimes and extinction dynamics.** **a**, Cheat-embracing (CE) and cheat-purging (CP) regimes. **b**, Each line is founded by a single WS genotype (yellow). After 6 days in phase I cells are harvested, plated and screened for SM colonies (red). Pooled SM colonies found phase II. After 3 days cells are harvested, plated and screened for WS colonies. A single WS colony of the dominant type founds the next generation. To avoid extinction, the WS mat of each line must be intact at the end of phase I and WS types must have produced SM cells; by the end of phase II, SM types must have produced WS cells. **c**, Persistence of lines ( $n = 120$ ) and causes of extinction.

among lineages might allow the possibility for life-cycle-enhancing mutations, which are beneficial over the longer time scale of the life cycle, to outrun life-cycle-disrupting mutations<sup>16,35,36</sup>.

### Adaptive evolution and fitness decoupling

To allow for selection among lines, we took 120 microcosms, each containing a single WS mat, and divided these into 15 replicate populations,



**Figure 2 | Evolution under the CE regime.** **a**, Lines ( $n = 120$ ) are arranged as 15 replicate populations of 8 lines each (one replicate is depicted). The experiment involved a total of 2,400 microcosms and 140 days of selection. Extinction events provide opportunity for viable lines to reproduce. Reproduction allows export of a viable line to a new microcosm. Lines marked for reproduction were chosen at random and replacements took place within the same replicate. **b**, Fitness of derived lines and their single cell constituents relative to ancestral (ANC) types: line fitness increased significantly, whereas cell-level fitness decreased. One lineage generation was performed for each of three replicate fitness assays for each line. Error bars are s.e.m., based on  $n = 14$ .

each composed of 8 lines (Fig. 2a). Lines that failed to complete the life cycle provided an opportunity for viable lines to export their success to a new microcosm. Upon the demise of a particular lineage, a viable line within the same replicate was chosen at random and allowed to replace the extinct type (Fig. 2a). Extinction occurred at high frequency resulting in ~5 replacements per generation (per replicate). After ten life-cycle generations each population housed viable lines. Selection on lineage viability—and concomitantly fecundity—was thus central to persistence.

To determine the course of evolution, the fitness of isolates from evolved lines was measured relative to ancestral types. A single WS genotype representative of each population (of eight lines) was taken at the end of the selection period (see Methods and Extended Data Fig. 1a). In addition, 15 independent WS genotypes were obtained one mutational step from the ancestral SM genotype, thus providing a 'baseline' reference for ancestral fitness (see Methods and Extended Data Fig. 1a). All representative WS genotypes were competed against a *lacZ*-marked reference strain (see Methods and Extended Data Fig. 1b, left panel), allowing the competitive performance of all ancestral and evolved types to be assessed against a single common genotype<sup>22</sup>. From a multi-level selection perspective<sup>37</sup>, any trait may have different impacts on the fitness of cycling lines and individual cells. To address this issue, we introduce two different fitness measures. Fitness of lines was defined as the number of WS mat offspring left relative to the marked competitor (Extended Data Fig. 1b, left panel); and cell fitness was assessed as the total number of cells contained within individual mats (irrespective of WS or SM type) at the end of the phase I period. Additional measures of cell-level performance are described below.

Fitness of evolved lines—as determined by the ability to leave mat offspring relative to ancestral types—improved significantly ( $\chi^2 = 4.262$ , degrees of freedom (d.f.) = 1,  $P = 0.039$ ; Fig. 2b and Extended Data Table 1a). This is consistent with an evolutionary response to selection and shows that when SM cells are integral to the reproduction of WS mats, lines not only persist, but their ecological performance improves.

While fitness of evolved lines improved, cell fitness significantly decreased ( $t_{79} = 3.092$ ,  $P = 0.0027$ ; Fig. 2b and Extended Data Table 1a). This is a notable result: success at the level of evolving lineages has come at a cost to the individual cells of which the lines are composed. The fact that fitness of the evolved lines is no longer explicable in terms of the fitness of individual cells indicates that lineage fitness has become decoupled from individual cell fitness. This is consistent with theoretical predictions that during major evolutionary transitions selection shifts

from the lower (cell) to the higher (collective) level<sup>2,38</sup>. With such a shift arises a new kind of biological individual whose emergence is likely to curtail the independent evolution of lower-level entities<sup>5</sup>.

### Phenotypic traits underpinning lineage improvement

Traits inherent in the individual cells must explain the improved reproductive capacity of lines<sup>37</sup>. To explore adaptations that contributed to increased fitness of lines, life history properties were determined relative to ancestral types. Three replicate microcosms of each representative genotype were destructively sampled each day throughout phase I and II and the frequency of each type determined. With interest in the possibility that selection might have tuned life history characteristics to suit the duration of each phase, the number of days of propagation was doubled in both phase I (12 days) and phase II (6 days).

Evolved lines increased their capacity to generate the phenotype required for the next stage of the life cycle. This was evident in both phases: the new type was produced earlier and more reliably compared to the ancestral baseline lineages ( $\chi^2 = 5.442$ , d.f. = 1,  $P = 0.0197$ ; Fig. 3a and Extended Data Table 1b). Moreover it was maximal at day six, suggesting tuning to the periodicity of the selection regime. Enhanced capacity to generate each stage of the life cycle was not explained by an increase in total cell density (density of the ancestral lineages was greater than that of the derived types  $F_1 = 51.521$ ,  $P < 0.0001$ ; Fig. 3b and Extended Data Table 1b).

The level of SM occurrence in phase I, and WS occurrence in phase II, is strongly related to line fitness (Extended Data Fig. 2a, b) in both the

derived and ancestral types (evolved:  $\chi^2 = 12.324$ , d.f. = 1,  $n = 14$ ,  $P = 0.0004$ ; baseline:  $\chi^2 = 22.801$ , d.f. = 1,  $n = 15$ ,  $P < 0.0001$ ; Extended Data Table 2). Notably, five ancestral lines with high fitness had a tendency towards early production of the next life phase of the life cycle (Extended Data Fig. 2a). After selection, this propensity for reliable production was present in the majority of lines (Extended Data Fig. 2b). In successful lines—those lines capable of completing the life cycle—this was not attributable to an increase in the number of cells in the next phase of the life cycle ( $F_1 = 0.589$ ,  $P = 0.4431$ ; Extended Data Fig. 3a and Extended Data Table 1b), nor to a change in the proportion of cells of the next phase ( $F_1 = 1.701$ ,  $P = 0.1926$ ; Extended Data Fig. 3b and Extended Data Table 1b). In fact the proportion of cells (and number of cells) marking the next stage of the life cycle was often higher in ancestral lineages (Extended Data Figs 3a, b).

The increase in the frequency of occurrence of the alternative cell type suggests that the transition rate between stages of the life cycle—a form of development—may have been the focus of selection. However, an alternative possibility is that the growth rate of cells has increased such that it is more likely for a given cell type (for example, SM) to reach critical numbers before the end of the phase in which it was generated (for example, phase I). Furthermore, higher cell growth could lead to a greater cumulative number of cell divisions and thus an increased chance of generating the next cell type (for example, WS during phase II), however the abundance patterns in Fig. 3b stand in contrast to this idea. A simple mathematical model shows that the possibilities for producing cells of the opposing type via evolution of increased rates of switching readily outpaces any such possibilities arising from improvement in cell growth rate<sup>31</sup>. Nonetheless, to see whether growth rate had changed we determined the maximum growth rate of SM cells.

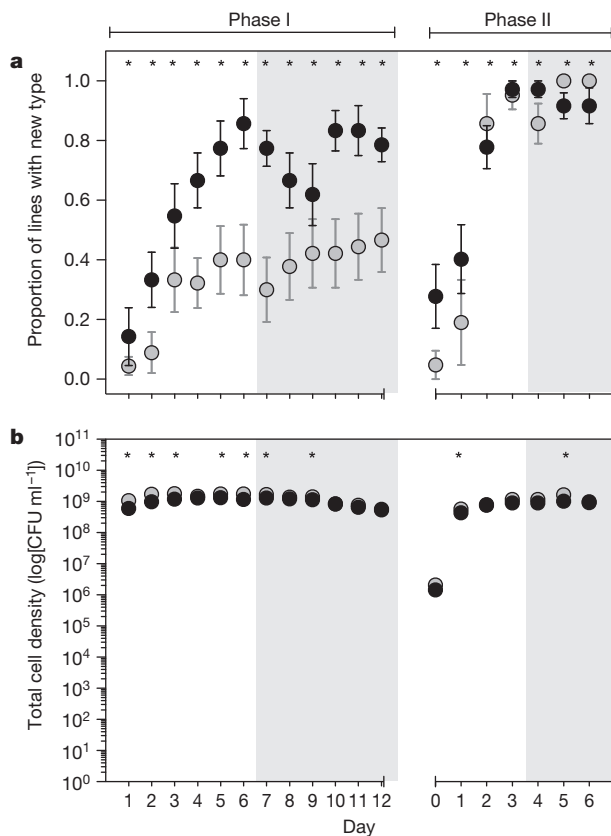
Compared to the growth rate of SM cells harvested from the ancestral lineages, the growth rate of SM cells harvested from the derived CE lines did not increase ( $t_7 = 1.527$ ,  $P = 0.1315$ ; Extended Data Fig. 4 and Extended Data Table 1a). This leaves increase in the rate of transition as a clear remaining explanation for the enhanced likelihood of detecting the new type in derived lines.

Natural selection can operate on a trait at any given hierarchical level provided the trait is heritable and covaries with fitness at that level, but determining the appropriate level is challenging, especially during transitions where selection is expected to act simultaneously at multiple levels<sup>2,39</sup>. To discern the level of selection at which traits are associated with fitness we performed a regression and correlation analysis on traits of lines and the single cells of which they are composed. The increased capacity to transition from WS to SM predicts line fitness in the derived lineages (Extended Data Figs 5a, b and Extended Data Table 2). Moreover, line fitness is not related to, and therefore cannot be explained by, the competitive performance of single cells (Extended Data Fig. 5b). Increased line fitness is therefore a probable product of selection on properties favouring transitions between life cycle phases—a developmental programme—rather than isolated success of cells within the phases.

### Genetic traits underpinning lineage improvement

The capacity for the fittest lines to transition through phases of the life cycle was markedly more rapid and reliable in derived lines compared to ancestral types (Extended Data Fig. 6). Notable was the phenotypic similarity between recurrences suggestive of a specific switch-like mechanism. The genome of the fittest lineage, line 17, was sequenced at generations 4 (WS<sub>8</sub>) and 11 (WS<sub>22</sub>) (Extended Data Fig. 7a). In addition, four independent replicate colonies of the 11-generation WS type (WS<sub>22</sub>) were transferred through three additional rounds of the two-phase life cycle culminating in four independent generation-14 WS types: the genome of one of these WS types (WS<sub>28</sub>) was sequenced, plus its immediate SM predecessor (SM<sub>27</sub>) (Extended Data Fig. 7a).

If the life cycle is initially driven by spontaneous mutation, then seven mutations should distinguish generation-4 WS (WS<sub>8</sub>) from ancestral SM (SM<sub>1</sub>). Comparative analysis revealed five mutations in loci known to effect expression of WS and SM types<sup>23</sup>. Next we interrogated the



**Figure 3 | Life history traits under the CE regime.** **a**, Proportion of lines producing the morphotype required for the next phase of the life cycle (that is, SM from WS during phase I and WS from SM during phase II). **b**, Total cell density. Black, derived; grey, ancestral. Grey shaded panels indicate the extension of each phase as compared to the selection regime. Error bars are s.e.m., based on  $n \leq 15$ . \* $P < 0.05$ , using a generalized linear model (error structure: binomial; link function: logit) and subsequent post hoc contrasts per day for **a**; and analysis of variance (ANOVA) and subsequent post hoc contrasts per day for **b**. CFU, colony forming unit.



genome of the generation-11 WS (WS<sub>22</sub>). This differed from the ancestral SM type by 53 mutations (Supplementary Table 1). This unexpectedly large number of mutations is a consequence of an elevated mutation rate attributable to a single nucleotide polymorphism in *mutS* (A1489C, which leads to a Tyr497Pro substitution).

An obvious question is whether the *mutS* (A1489C) mutation has a direct role in promoting switching. An alternative possibility would be that *mutS* increased the probability of a different mutation that directly caused elevated switching and then hitchhiked with the switch-causing mutation. The mutant *mutS* allele was reverted to wild type and the capacity of line-17 wild-type *mutS* (*mutS*<sup>WT</sup>) to pass rapidly and repeatedly through phases of the life cycle was checked. The *mutS*<sup>WT</sup> line was significantly impaired in this capacity, indicating that switching depends directly on *mutS*.

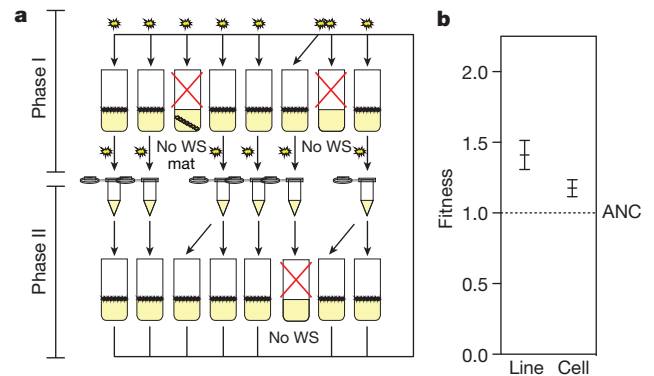
The genome of a single generation-14 WS (WS<sub>28</sub>) was next compared to its immediate SM (SM<sub>27</sub>) progenitor (Extended Data Fig. 7a and Supplementary Table 1). Because of the mutator background, several mutations had fixed in the WS genome; however, of note was a frameshift mutation in a tract of seven guanine residues in *wspR* (nucleotides 742–748) resulting in an additional guanine in the SM type. *wspR* encodes a diguanylate cyclase and is one of 39 such genes that can, in principle, underpin expression of the WS phenotype (and its loss)<sup>23,26,27</sup>. Given that the tract of guanine residues overlaps the active site of WspR (ref. 40) we considered the possibility that the *mutS* mutation might have turned *wspR* into a genetic switch enabling rapid and predictable transitioning between SM and WS through expansion and contraction of the tract of guanines<sup>41</sup> in a manner reminiscent of a contingency locus<sup>42</sup>. A disproportionate increase in the number of frameshift mutations in the *mutS* genotype, specifically in tracts of guanine (and corresponding cytosine) residues (Supplementary Table 1), supports this possibility<sup>43</sup>.

The DNA sequence of *wspR* was obtained from every WS and SM genotype that arose from fourfold replication of the generation-11 line-17 WS through three additional ‘expedited’ rounds of the life cycle (Extended Data Fig. 7a). As shown in Extended Data Fig. 7b many of the transitions between WS and SM from generation 10 onward correlated with expansion (*wspR* OFF) or contraction (*wspR* ON) of the guanine tract. In a control experiment, performed with line-17 *mutS*<sup>WT</sup>, the slower and less reliable transition between phenotypic states did not, with a single exception, involve the tract of guanine residues.

The existence of a genetic switch in line-17 is highly advantageous to the collective: it strengthens heritability between recurrences; integrates both phases into essentially a single entity; and constitutes a critical first step in the emergence of differentiation<sup>4</sup>. Genome sequencing showed that the *mutS*-dependent switch arose in just a single lineage, but fixed in all eight populations of the replicate. Its evolution was reliant on earlier mutations in *Wsp* that preserved functionality of the pathway while ensuring constitutive activation of WspR<sup>25</sup>. It also depended on mutations elsewhere in the genome that exhausted alternate genetic routes to WS. Together, this set of prior mutations, in conjunction with *mutS* (A1489C), conferred special significance to the tract of guanine residues in *wspR*. While dependency on *mutS* might seem a dangerous liaison, the life cycle provides ample opportunity for purifying selection to maintain integrity of the SM type and refine the switch.

## Purging cheats

Reproduction of mats in the cheat-embracing (CE) regime underpinned significant evolutionary change. The causative factor is of central interest. While it is possible that a life cycle of two phases is key, reproduction via a bottleneck phase, combined with selection among lineages, may be sufficient. To test for such a possibility we performed a control experiment in which WS mats were propagated under a ‘cheat-purging’ (CP) regime. The sole difference between the CP and CE regimes is the cell phenotype that passes through the bottleneck: under the CP regime the cell passing through the bottleneck is a WS cell (Fig. 1a, right panel). Reproduction via the CP regime is analogous to fragmentation.



**Figure 4 | Evolution under the CP regime.** **a**, Lines ( $n = 120$ ) are arranged as 15 replicates of eight (one replicate is depicted). The experiment involved a total of 2,400 microcosms. Extinction events provide an opportunity for viable lines to reproduce. Reproduction allows export of a viable line to a new microcosm. Lines marked for reproduction were chosen at random and replacements took place within the same replicate. **b**, Fitness of derived lines and their single cell constituents relative to ancestral (ANC) types: both line and cell-level fitness increased significantly. One lineage generation was performed for each of three replicate fitness assays for each line. Error bars are s.e.m., based on  $n = 15$ .

Conventional wisdom predicts that the CP regime will have the greatest evolutionary potential. Indeed, after 10 generations of lineage selection (Fig. 4a) fitness of evolved lines improved significantly ( $\chi^2 = 15.737$ , d.f. = 1,  $P < 0.0001$ ; Fig. 4b and Extended Data Table 1a). However, this was accompanied by a similar improvement in the fitness of single cells ( $t_{86} = 2.132$ ,  $P = 0.036$ ; Fig. 4b and Extended Data Table 1a). Improvement in the fitness of evolved lines can be explained solely by improvement in individual cell performance. The striking response observed under the CE regime can therefore be attributed to a life cycle of alternating phases.

To explore adaptations of the CP regime that contributed to increased lineage performance, life history properties were determined relative to ancestral types, as for the CE regime. No increase was seen in the capacity for WS types to transition to SM. Indeed, a significantly lower proportion of the derived groups produced SM ( $\chi^2 = 8.199$ , d.f. = 1,  $P = 0.0042$ ) and SM types took longer to arise (Extended Data Fig. 8a), hinting at the possibility that cheat suppression might have begun to evolve under this regime (Extended Data Figs 8b–d).

Under the CP regime enhanced line fitness in the derived lineages is explained by changes in traits that improve the competitive ability of individual cells (Extended Data Figs 5c, d and Extended Data Table 2). Enhanced line fitness under the CP regime can be viewed as a by-product of selection at the lower (cell) level. The rate of transition between WS and SM correlates negatively with cell fitness parameters and is not associated with the CP-line fitness. In contrast to the CE regime, in which the WS to SM transition rate increased in a manner interpretable as a consequence of selection at the level of collectives, this trait decreased in the CP regime consistent with selection operating at the cell level. Evolution of this trait in opposing directions can be viewed as resulting from selection at different levels.

## Perspective

Multicellular organisms are descendants of once free-living cells<sup>1,3,4</sup>. By virtue of their capacity for differential reproduction, ancestral free-living cells were units of selection<sup>33</sup>. During the transition to multicellularity, collectives of cells emerged that came to participate in Darwinian processes in their own right<sup>2,5,11,19</sup>. The essential ingredient was a means of collective reproduction<sup>5,11</sup>. This most seminal of Darwinian properties emerges afresh at each transition and requires explanation<sup>44</sup>. Here we have shown that cheating cells—those types seemingly most detrimental to the persistence of newly formed cooperative entities—can

function as a germ line within a life cycle that facilitates the reproduction of collectives. Moreover, the two-phase life cycle presents selection with an altogether new kind of biological entity: each state becomes a different attribute of a single organism whose evolution is unified through a developmental programme<sup>45</sup>. When reproduction of collectives is via fragmentation (a single-phase life cycle), the traits that yield success at the higher level are largely those that determine success of single cells. This offers limited opportunity for the emergence of new kinds of biological individuality because properties of higher and lower levels remain aligned<sup>19,38</sup>.

Direct observation of early stages in an evolutionary transition requires that issues surrounding levels of selection be considered<sup>2,5,19</sup>. This necessarily leads to territory in which a range of perspectives is possible (see Supplementary Discussion). Our experimental design incorporates an ecology that is explicitly multi-level: both individual cells (that reproduce once every hour), and individual lineages (that reproduce once every 9 days) can be units of selection; however, selection operates on cells and lineages over different timescales. While selection on individual cells favours short-term success, short-term success is unlikely to facilitate persistence of lineages. Indeed, persistence requires more than simply switching between phenotypes: it involves a developmental programme that underpins expression of a collective phase in which a soma-like body is constructed from germ-like cells. Cells of the body must simultaneously play an ecological role (maintaining the body near oxygen via a robust mat phenotype) while producing the seeds of the next generation of bodies (the germ-like cells). Given sufficient variation among lineages, then selection over the longer timescale stands to conquer the short-term interests of individual cells. This appears to have happened in our CE regime with decoupling of fitness between levels supporting the view that selection has begun the process of transitioning to the higher (collective) level—with the lower level beginning to function for the good of the collective.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 3 November 2013; accepted 22 September 2014.**

- Bonner, J. T. The origins of multicellularity. *Integr. Biol.* **1**, 27–36 (1998).
- Okasha, S. *Evolution and the Levels of Selection* (Oxford Univ. Press, 2006).
- Maynard Smith, J. & Szathmari, E. *The Major Transitions in Evolution* (Oxford Univ. Press, 1995).
- Buss, L. W. *The Evolution of Individuality* (Princeton Univ. Press, 1987).
- Godfrey-Smith, P. *Darwinian Populations and Natural Selection* (Oxford Univ. Press, 2009).
- Nowak, M. A. Five rules for the evolution of cooperation. *Science* **314**, 1560–1563 (2006).
- Sachs, J. L., Mueller, U. G., Wilcox, T. P. & Bull, J. J. The evolution of cooperation. *Q. Rev. Biol.* **79**, 135–160 (2004).
- Rainey, P. B. & De Monte, S. Resolving conflicts during the evolutionary transition to multicellular life. *Annu. Rev. Ecol. Syst.* **45**, 599–620 (2014).
- Rainey, P. B. & Rainey, K. Evolution of cooperation and conflict in experimental bacterial populations. *Nature* **425**, 72–74 (2003).
- Pfeiffer, T., Schuster, S. & Bonhoeffer, S. Cooperation and competition in the evolution of ATP-producing pathways. *Science* **292**, 504–507 (2001).
- Maynard Smith, J. in *Evolutionary Progress* (ed. Nitecki, M. H.) 219–230. (Univ. Chicago Press, 1988).
- Sober, E. & Wilson, D. S. *Unto Others: The Evolution and Psychology of Unselfish Behaviour* (Harvard Univ. Press, 1998).
- Velicer, G. J., Kroos, L. & Lenski, R. E. Developmental cheating in the social bacterium *Myxococcus xanthus*. *Nature* **404**, 598–601 (2000).
- Strassmann, J. E., Zhu, Y. & Queller, D. C. Altruism and social cheating in the social amoeba *Dictyostelium discoideum*. *Nature* **408**, 965–967 (2000).
- Michod, R. E. Cooperation and conflict in the evolution of individuality. 2. Conflict mediation. *Proc. R. Soc. Lond. B* **263**, 813–822 (1996).
- Nunney, L. Group selection, altruism, and structured-deme models. *Am. Nat.* **126**, 212–230 (1985).
- Wade, M. J. & Breden, F. The evolution of cheating and selfish behavior. *Behav. Ecol. Sociobiol.* **7**, 167–172 (1980).
- Santorelli, L. A. et al. Facultative cheater mutants reveal the genetic complexity of cooperation in social amoebae. *Nature* **451**, 1107–1110 (2008).
- Michod, R. E. *Darwinian Dynamics: Evolutionary Transitions in Fitness and Individuality* (Princeton Univ. Press, 1999).
- Frank, S. A. Mutual policing and repression of competition in the evolution of cooperative groups. *Nature* **377**, 520–522 (1995).
- Queller, D. C. Relatedness and the fraternal major transitions. *Phil. Trans. R. Soc. Lond. B* **355**, 1647–1655 (2000).
- Rainey, P. B. Unity from conflict. *Nature* **446**, 616 (2007).
- McDonald, M. J., Gehrig, S. M., Meintjes, P. L., Zhang, X. X. & Rainey, P. B. Adaptive divergence in experimental populations of *Pseudomonas fluorescens*. IV. Genetic constraints guide evolutionary trajectories in a parallel adaptive radiation. *Genetics* **183**, 1041–1053 (2009).
- Rainey, P. B. & Travisano, M. Adaptive radiation in a heterogeneous environment. *Nature* **394**, 69–72 (1998).
- Bantinaki, E. et al. Adaptive divergence in experimental populations of *Pseudomonas fluorescens*. III. Mutational origins of wrinkly spreader diversity. *Genetics* **176**, 441–453 (2007).
- Goymer, P. et al. Adaptive divergence in experimental populations of *Pseudomonas fluorescens*. II. Role of the GGDEF regulator WspR in evolution and development of the wrinkly spreader phenotype. *Genetics* **173**, 515–526 (2006).
- Spier, A. J., Kahn, S. G., Bohannon, J., Travisano, M. & Rainey, P. B. Adaptive divergence in experimental populations of *Pseudomonas fluorescens*. I. Genetic and phenotypic bases of wrinkly spreader fitness. *Genetics* **161**, 33–46 (2002).
- Tarnita, C. E., Taubes, C. H. & Nowak, M. A. Evolutionary construction by staying together and coming together. *J. Theor. Biol.* **320**, 10–22 (2013).
- Trivers, R. L. The evolution of reciprocal altruism. *Q. Rev. Biol.* **46**, 35–57 (1971).
- Rainey, P. B. & Kerr, B. Cheats as first propagules: a new hypothesis for the evolution of individuality during the transition from single cells to multicellularity. *Bioessays* **32**, 872–880 (2010).
- Libby, E. & Rainey, P. B. A conceptual framework for the evolutionary origins of multicellularity. *Phys. Biol.* **10**, 035001 (2013).
- Libby, E. & Rainey, P. B. Eco-evolutionary feedback and the tuning of proto-developmental life cycles. *PLoS ONE* **8**, e82274 (2013).
- Lewontin, R. C. The units of selection. *Annu. Rev. Ecol. Syst.* **1**, 1–18 (1970).
- Wolpert, L. & Szathmari, E. Multicellularity: evolution and the egg. *Nature* **420**, 745 (2002).
- Leigh, E. G. How does selection reconcile individual advantage with the good of the group? *Proc. Natl Acad. Sci. USA* **74**, 4542–4546 (1977).
- Nunney, L. Lineage selection and the evolution of multistage carcinogenesis. *Proc. R. Soc. Lond. B* **266**, 493–498 (1999).
- Damuth, J. & Heisler, I. L. Alternative formulations of multi-level selection. *Biol. Philos.* **3**, 407–430 (1988).
- Michod, R. E. & Roze, D. in *Mathematical and Computational Biology: Computational Morphogenesis, Hierarchical Complexity, and Digital Evolution* (ed. Nehaniv, C. L.) 47–92 (American Mathematical Society, 1999).
- Okasha, S. Emergence, hierarchy and top-down causation in evolutionary biology. *Interface Focus* **2**, 49–54 (2012).
- De, N. et al. Phosphorylation-independent regulation of the diguanylate cyclase WspR. *PLoS Biol.* **6**, e67 (2008).
- Levinson, G. & Gutman, G. A. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol. Biol. Evol.* **4**, 203–221 (1987).
- Moxon, E. R., Rainey, P. B., Nowak, M. A. & Lenski, R. E. Adaptive evolution of highly mutable loci in pathogenic bacteria. *Curr. Biol.* **4**, 24–33 (1994).
- Heilbron, K., Toll-Riera, M., Kojadinovic, M. & MacLean, R. C. Fitness is strongly influenced by rare mutations of large effect in a microbial mutation accumulation experiment. *Genetics* **197**, 981–990 (2014).
- Griesemer, J. The units of evolutionary transition. *Selection* **1**, 67–80 (2001).
- Wolpert, L. The evolution of development. *Biol. J. Linn. Soc.* **39**, 109–124 (1990).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** We thank S. Nestmann for assistance with molecular aspects of the work and for guiding construction of the *mutS* deletion mutant. We thank E. Libby and Y. Pichugin for discussion, and S. De Monte and P. G. Smith for comments on drafts of the manuscript. We are indebted to PacBio and particularly J. Korlach and Y. Song for genome sequencing. P.B.R. currently holds an International Blaise Pascal Research Chair funded by the French State and the Ile-de-France, managed by the Fondation de l'Ecole Normale Supérieure. The work was directly supported by the Marsden Fund Council from government funding administered by the Royal Society of New Zealand, and in part by grant RFP-12-20 from the Foundational Questions in Evolutionary Biology Fund, by the National Science Foundation under Cooperative Agreement Number DBI-0939454, and by an NSF CAREER Award Grant (DEB0952825).

**Author Contributions** All authors contributed to the conception and design of the study. K.H. and C.J.R. performed research, undertook data analysis and prepared figures. All authors wrote the paper.

**Author Information** All genome data have been deposited into the Sequence Read Archive under accession number SRP047104. P.B.R. will make strains available to qualified recipients. Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to P.B.R. (p.b.rainey@massey.ac.nz).



## METHODS

**Strains and medium.** *Pseudomonas fluorescens* SBW25 (ref. 46) was grown at 28 °C in 25-ml static glass microcosms containing 6 ml of King's Medium B (with loose caps), and on King's Medium B agar plates for 48 h. For competitive assays, strains marked with *lacZ* (ref. 47) allowed types to be distinguished by plating on media containing 60 µg ml<sup>-1</sup> X-gal.

**CE regime.** The CE regime (Fig. 1a, left panel) involved a two-phase life cycle (Fig. 1b). The experiment was founded by a single ancestral SBW25 genotype that initiated phase II. After 3 days of static incubation (phase II), 1.5 × 10<sup>-7</sup> ml (50 µl of a 3.35 × 10<sup>5</sup>-fold dilution) was plated and a single WS colony of the most abundant morphotype was picked from each plate to inoculate a fresh microcosm (Fig. 1b). This marked the start of phase I of the first generation. Following 6 days of static incubation (phase I), microcosms were visually inspected to check for the presence of an intact mat. Lines that had no mat at day six were deemed extinct. Microcosms containing lines with intact mats were vortex mixed and 2.5 × 10<sup>-8</sup> ml (200 µl of a 8.0 × 10<sup>6</sup>-fold dilution) plated on solid media. After 48 h incubation plates were inspected. To avoid extinction and successfully passage to phase II it was necessary for lines to have produced SM types. All SM colonies were transferred to 200 µl liquid medium and incubated for 24 h under static conditions. The pooled set (from each microcosm) were mixed and 6 µl was used to inoculate phase II microcosms.

**CP regime.** The CP regime (Fig. 1a, right panel) was identical to the CE regime except that during the life cycle cheats were purged and WS founded both phases of the cycle. Both regimes were conducted in parallel.

**Between-lineage selection.** Each treatment consisted of 15 replicates of 8 competing microcosms, and replicates from both treatments were spread evenly across 4 experimental blocks. All microcosms in the CE regime were subjected to the selection regime outlined above; however, following assessment of mat integrity at the end of phase I, surviving lineages within each replicate were harvested by vortex-mixing and dilutions spread on agar plates (Fig. 2a). Extinct lineages (due to mat collapse/no mat/no SM colony post phase I, or no WS colony post phase II) were immediately replaced by randomly chosen surviving lines taken from the same replicate. On rare occasions all eight lines of a replicate were eliminated: in these instances one line, chosen at random, from the same experimental block, was used to re-found the replicate. The CP treatment was carried out as above, differing only following phase I when WS colonies were picked instead of SM colonies (Fig. 4a). A smaller volume (6.25 × 10<sup>-9</sup> ml (50 µl of a 8.0 × 10<sup>6</sup>-fold dilution)) was plated due to the absence of a phenotype-switching requirement for the CP treatment. Both treatment regimes were implemented for ten generations.

**Selection of representative WS genotypes.** One single WS genotype to represent each replicate from the derived CE and CP 'between-lineage selection regimes' was generated after ten generations as described in Extended Data Fig. 1. The representative WS colonies were grown in shaken microcosms (16 h) and stored at -80 °C for post-selection analyses. This yielded 15 such types each for the ancestral and CP regimes, and 14 types for the CE regime (bacteria from one CE line could not be revived from the freezer stock).

A *lacZ*-marked reference strain (W1-*lacZ*) for fitness comparisons in the CP regime (Extended Data Fig. 1b) was generated the same way from 3-day static microcosms inoculated with a *lacZ*-marked ancestral strain SBW25-*lacZ* (ref. 47).

**Fitness assay.** Lineage- and cell-level fitness was assessed for all 44 representative ancestral and derived types (Extended Data Fig. 1b). The representative types were not directly competed against each other but against a single, neutrally marked reference strain assayed under the appropriate regime: the ancestral and derived CE types were competed against SBW25-*lacZ*, and the ancestral and derived CP types were competed against W1-*lacZ*. This yielded independent fitness values for each type (that is, potential interactions between ancestral and derived types didn't affect the result), and allowed for estimates and comparison of relative performance for all types.

One lineage generation was performed for each of three replicate fitness assays for all 44 representative types (as shown in Extended Data Fig. 1b for one CE and one CP replicate). For each representative type, eight microcosms were each inoculated with one WS colony in phase I. After mat assessment surviving microcosms were pooled, and 2 × 10<sup>-7</sup> ml (200 µl of a 1.0 × 10<sup>6</sup>-fold dilution) (CE) or 5 × 10<sup>-8</sup> ml (50 µl of a 1 × 10<sup>6</sup>-fold dilution) (CP) of this mixture plated. The total number of colonies (in mats) was recorded as a measure of cell fitness. All SM (CE) and WS (CP) colonies, and colonies of the competitor strains SBW25-*lacZ* and W1-*lacZ* were grown overnight and subsequently pooled as described above (CE regime). The derived and ancestral types were mixed with the competitor strains in proportion to their performance during phase I (details displayed in Extended Data Fig. 1b). Six microlitres of this mixture was used to inoculate the eight phase II microcosms. After three days of static incubation (phase II), microcosms were plated. The most abundant WS colony morphotype on each plate determined whether the ancestral/derived representative type or the marked reference type was more successful. The fitness of types representative of each line was calculated for each replicate

of eight microcosms as the number of successful offspring of this type as a proportion of the total number of potential offspring (Extended Data Fig. 1b).

**Life-history analysis.** Static microcosms (36 per representative genotype) were individually inoculated with single colonies of the representative WS types (1,584 microcosms in total). Each day, three replicates were destructively harvested, plated, and the number of SM and WS colony forming units per microcosm was recorded. phase I was extended from 6 to 12 days, phase II from 3 to 6 days. At day six, propagules were collected for phase II, and microcosms inoculated (18 per type). Each day, three replicate microcosms per representative type were destructively harvested and number of SM and WS colony forming units recorded.

**SM growth assay.** Three biological replicate SM colonies were derived from each of the representative ancestral and derived CE and CP WS types. Four day static microcosms seeded from a single WS colony were destructively harvested and plated (5 × 10<sup>-8</sup> ml). SM colonies were chosen, grown in shaken microcosms (16 h), and stored at -80 °C. This procedure was repeated until three biological SM replicate colonies were obtained. For the types where three biological replicate SM colonies couldn't be derived, additional experimental replicates of SM growth rate were assessed. In a small number of instances no SM types were obtained.

SM growth kinetics were determined in 96-well microtitre plates shaken at 28 °C, and absorbance (OD<sub>600</sub>) measured in a microplate reader (BioTek). Each well was inoculated with approximately 10<sup>4</sup> SM cells in 180 µl King's medium B and absorbance measured every 10 min for 24 h. The growth of each biological replicate was determined in three different well locations on independent 96-well plates and on separate days. The maximum growth rate ( $V_{max}$ ) was calculated from the maximum slope of the absorbance over time. The mean  $V_{max}$  for each representative type was calculated from all biological and experimental SM replicates.

**Statistical analysis.** For detecting differences in line level fitness between the ancestral and derived regimes, a generalized linear model (error structure: binomial; link function: logit) with the explanatory variables regime, and representative type (nested within regime) was calculated. Contrasts revealed differences in lineage level fitness between regimes.

Analysis of variance (ANOVA) was used to test for differences in total number of cells, number of WS per µl, number of SM per µl (if present), and SM growth rate between the different regimes. Explanatory variables were regime, and representative type (nested within regime). Post hoc contrasts revealed differences between the ancestral and their respective derived regime.

Generalized linear models (error structure: binomial; link function: logit) were used to test for the difference between regimes in life-history parameters during the course of the experiment. The response variable was 'proportion of microcosms with the new type'. Explanatory variables were regime, representative type (nested within regime), and time. Analysis of variance was performed with the same explanatory variables but for 'new cell type per µl', 'total cells per µl', and 'proportion of the new cell type within a microcosm'. All three variables were Box-Cox transformed. Contrasts revealed differences between the regimes on the individual days.

Relationships between all parameters were tested using the mean per representative type accounting for regime. Pearson and Spearman rank correlations, and regressions (line level fitness: generalized linear models with the normal error structure, and the identity link function; cell-level fitness and number of SM per µl (if present): general linear models) were performed. SM growth rate, number of SM per µl (if present), and SM/WS occurrence were Box-Cox transformed.

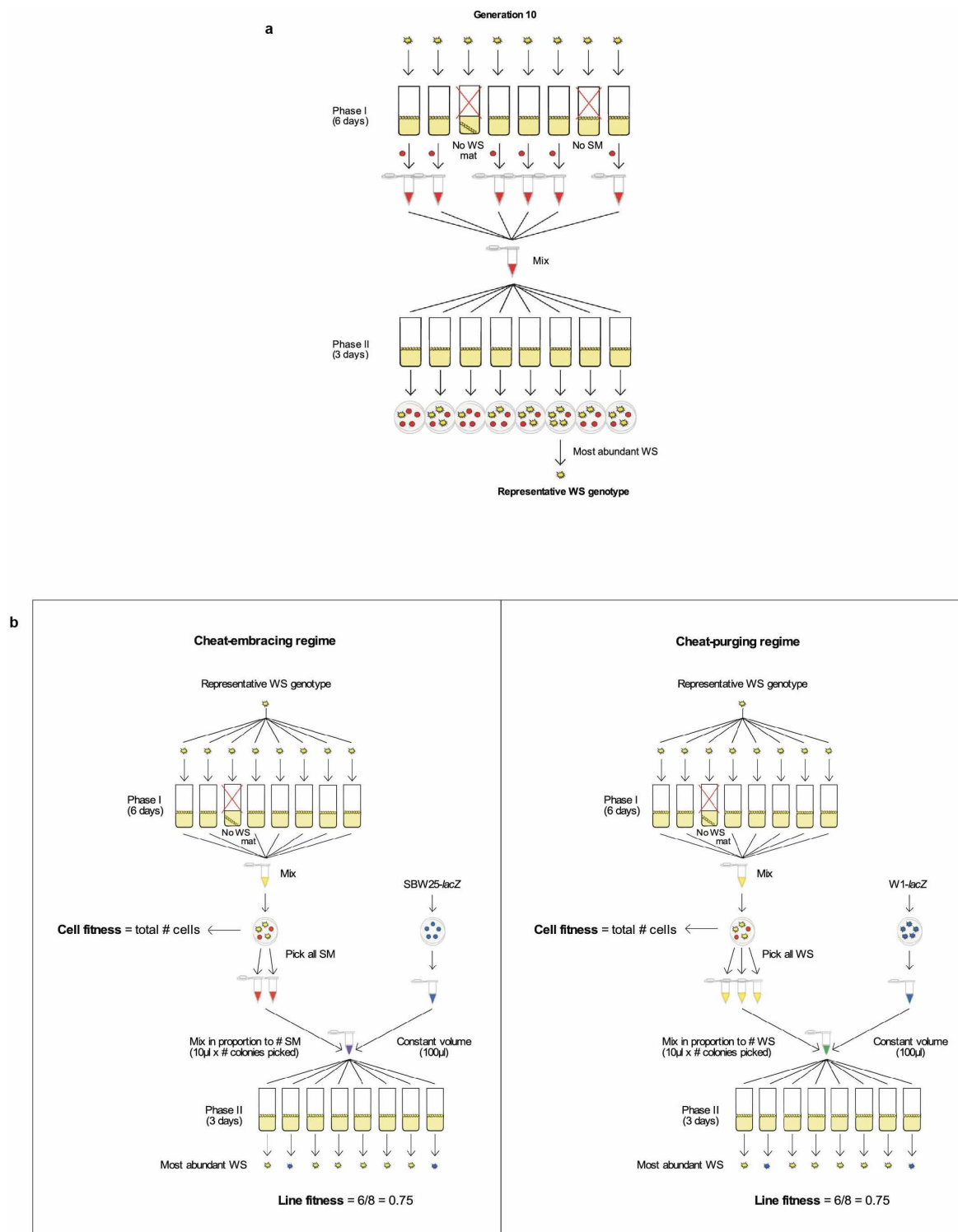
Colony counts from contaminated plates were excluded from analyses. Sample size was chosen to maximise statistical power and ensure sufficient replication. Assumptions of the tests, that is, normality and equal distribution of variances, were visually evaluated. Non-significant interactions were removed from the models. All tests were two-tailed. Effects were considered significant at the level of  $P < 0.05$ . All statistical analyses were performed with JMP 9. Graphs were produced with GraphPad Prism 5.0, JColorgrid (<http://jcolorgrid.sourceforge.net>), Adobe Illustrator CC 17.0.0 and Inkscape 0.48.2.

**Genetic manipulation.** Standard genetic techniques were used to revert *mutS* (A1489C) to wild type. This involved PCR amplification of the wild-type sequence from SBW25 and its integration into the genome of WS<sub>22</sub> by homologous recombination facilitated by a two-step allelic replacement strategy using pUIC-3 (ref. 27). Characterization of G-tract expansion and contraction in *wspR* was performed by PCR and Sanger sequencing.

**Genome sequencing.** Genomic DNA was extracted from overnight cultures (each founded by a single colony) using Wizard Genomic DNA Purification Kit (Promega), and sent to the Australian Genome Research Facility and Pacific Biosciences, for Illumina and SMRT sequencing, respectively. Libraries for SMRT sequencing were prepared on a Sciclone NGS automated liquid handling workstation (Perkin Elmer, Waltham MA) using a 10-kb automated library preparation protocol (Pacific Biosciences, Menlo Park, CA) which includes an initial Solid Phase Reversible Immobilization (SPRI) clean-up step, followed by the standard 10-kb library preparation protocol. Libraries were annealed using 20× excess primer using the standard annealing

protocol (Pacific Biosciences). Sequencing was performed on the PacBio RSII using P4-C2 sequencing chemistry, magnetic bead loading, and 3-h movie acquisitions. Sequence reads were mapped against the *P. fluorescens* SBW25 GenBank reference and variants were called using SMRTPortal version 2.3. Additional analyses were performed with Geneious 7.1.4.

46. Silby, M. W. *et al.* Genomic and genetic analyses of diversity and plant interactions of *Pseudomonas fluorescens*. *Genome Biol.* **10**, R51 (2009).
47. Zhang, X. X. & Rainey, P. B. Construction and validation of a neutrally-marked strain of *Pseudomonas fluorescens* SBW25. *J. Microbiol. Methods* **71**, 78–81 (2007).

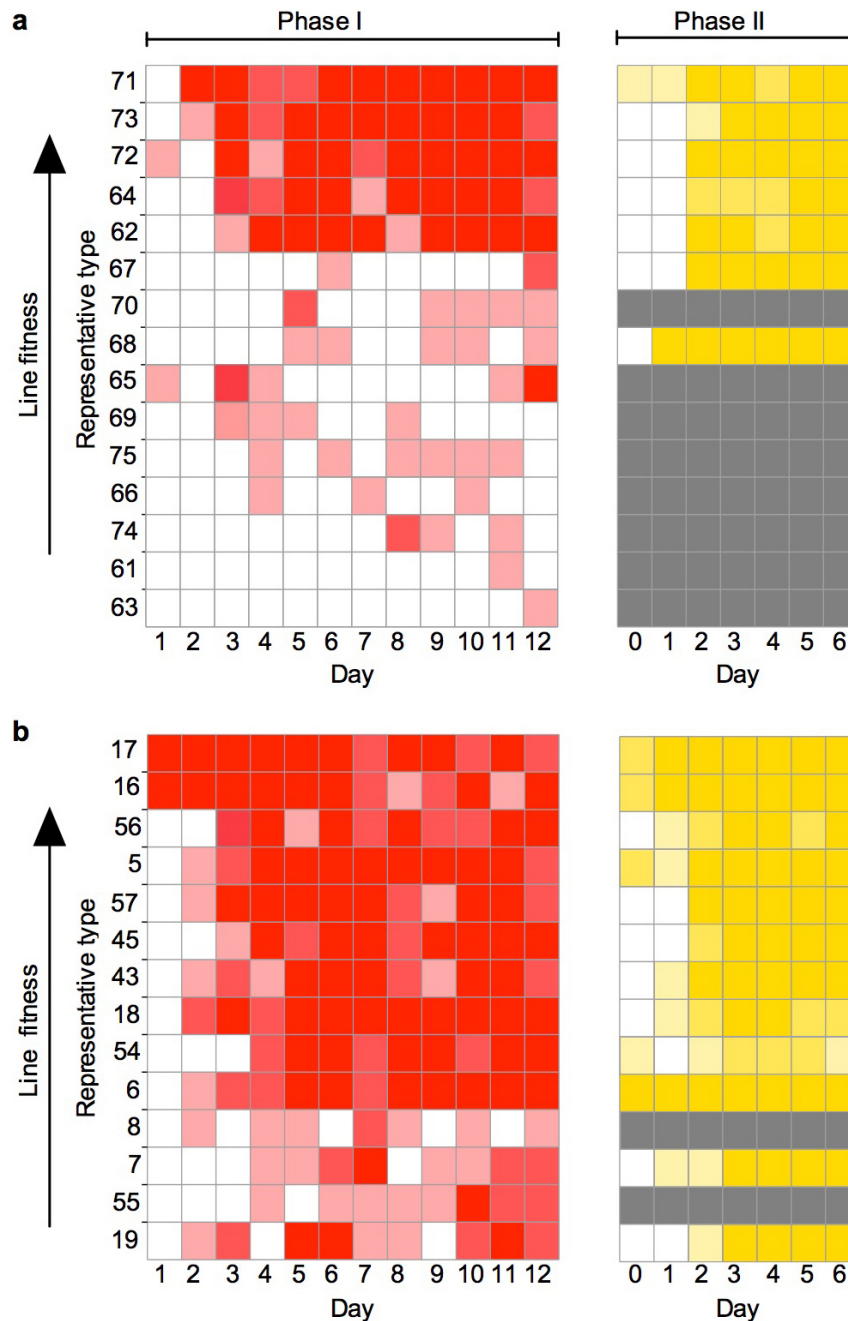


### Extended Data Figure 1 | Representative WS genotypes and fitness assays.

**a.** To analyse the response of derived lines to selection, a single representative WS genotype was obtained from each replicate population (set of eight microcosms) from both ancestral and derived lines. To obtain the representative set of derived types under the CE regime, SM colonies were collected from the end of phase I at generation 10 and pooled. The pooled sample was used to found phase II, at the end of which a single WS-type representative of each replicate was selected as described for the baseline (see below). This yielded 14 such types, one representing each replicate. To obtain the set of baseline types representing the ancestral state, SBW25 was used to found phase II. At the end of the 3-day period lines were harvested and plated. The single most abundant WS type from the most densely populated plate was selected as representative of that replicate. A third set of representative WS

genotypes was obtained from lines evolved under the CP regime. This was as for the CE regime, but instead of pooling SM at the end of phase I, WS were collected and pooled. A single dominant WS type was chosen from each replicate. **b.** Cell- and line-fitness assays. Lines founded by representative WS genotypes (from ancestral and derived lineages) were competed against a marked (blue colonies) reference strain (SM and WS, for the CE and CP regimes, respectively). Use of the marked reference strain allowed the competitive performance of all ancestral and derived types to be assessed against a single common genotype. Cell fitness is the total number of cells in the mat after phase I, whereas line fitness is the proportion of evolved 'offspring' mats relative to a marked reference strain. In total 2,472 microcosms were assayed (three replicate assays per line).

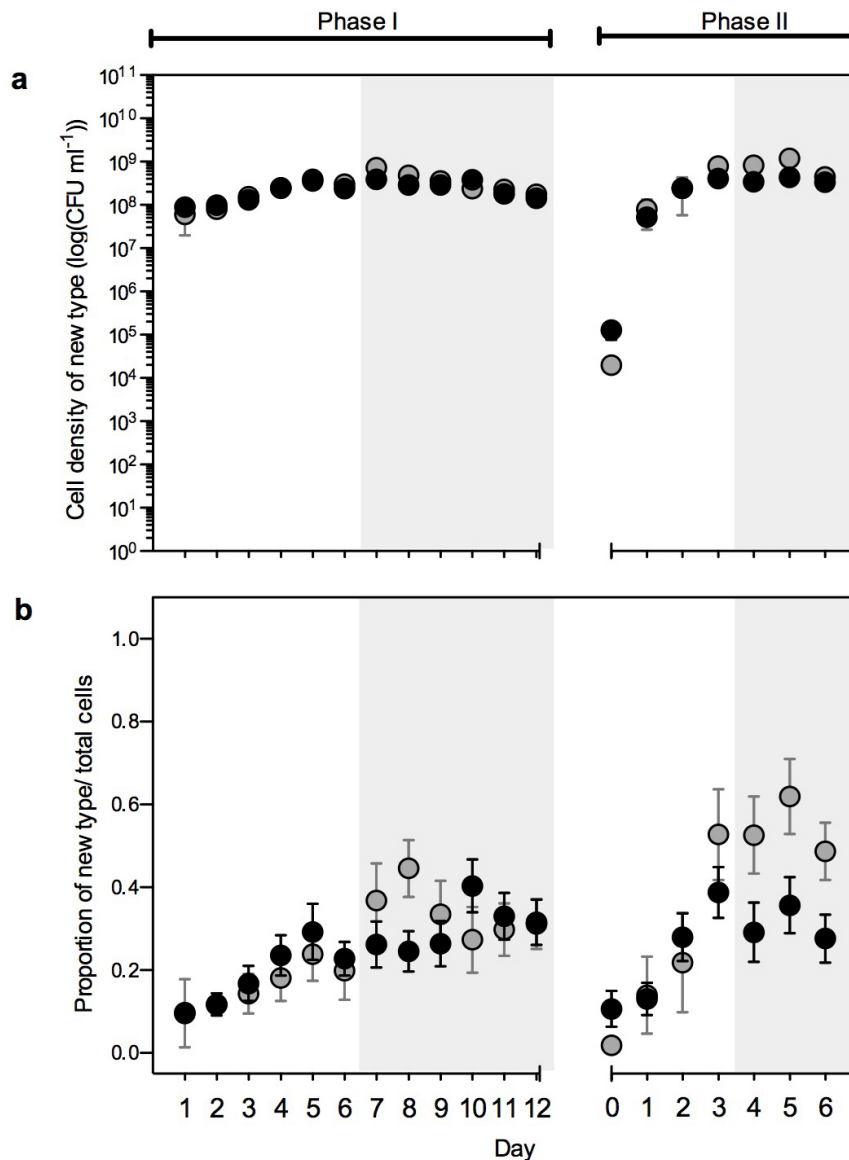




#### Extended Data Figure 2 | Line fitness and life cycle perpetuation.

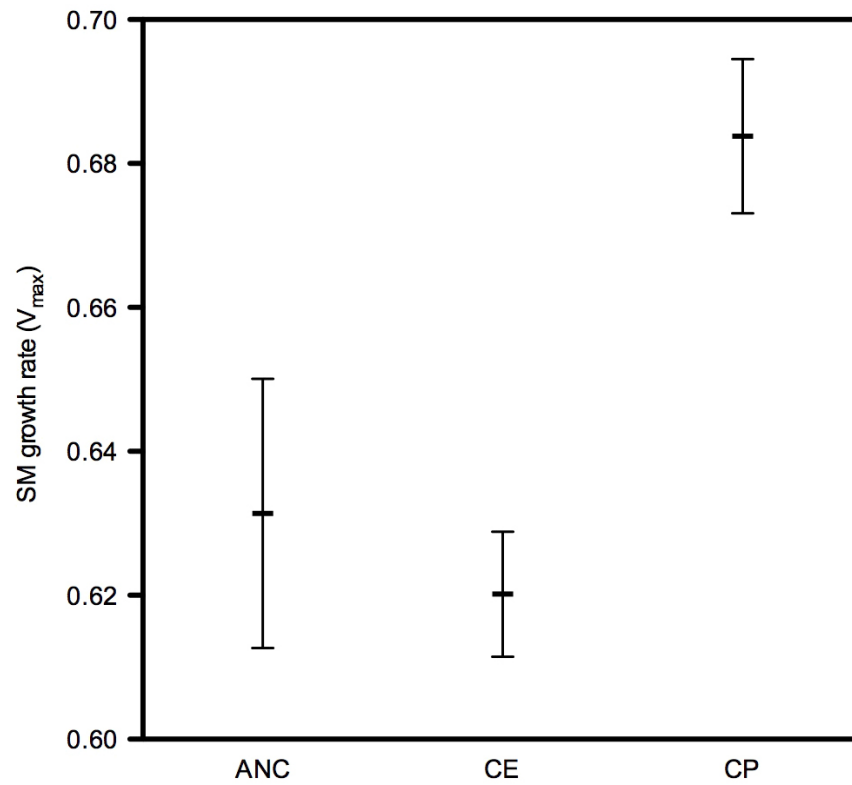
**a, b**, Fitness of ancestral (**a**) and derived (**b**) lines and relationship with capacity to perpetuate the two-phase life cycle. Data are a breakdown of data in Fig. 3a. Colour intensity represents the proportion of three replicate microcosms

harbouring the new type (white, absence of new type; dark red (yellow), presence of SM (WS) in all microcosms). Lines are ordered according to their fitness as assayed under the CE regime. Grey cells indicate loss of lines due to extinction.



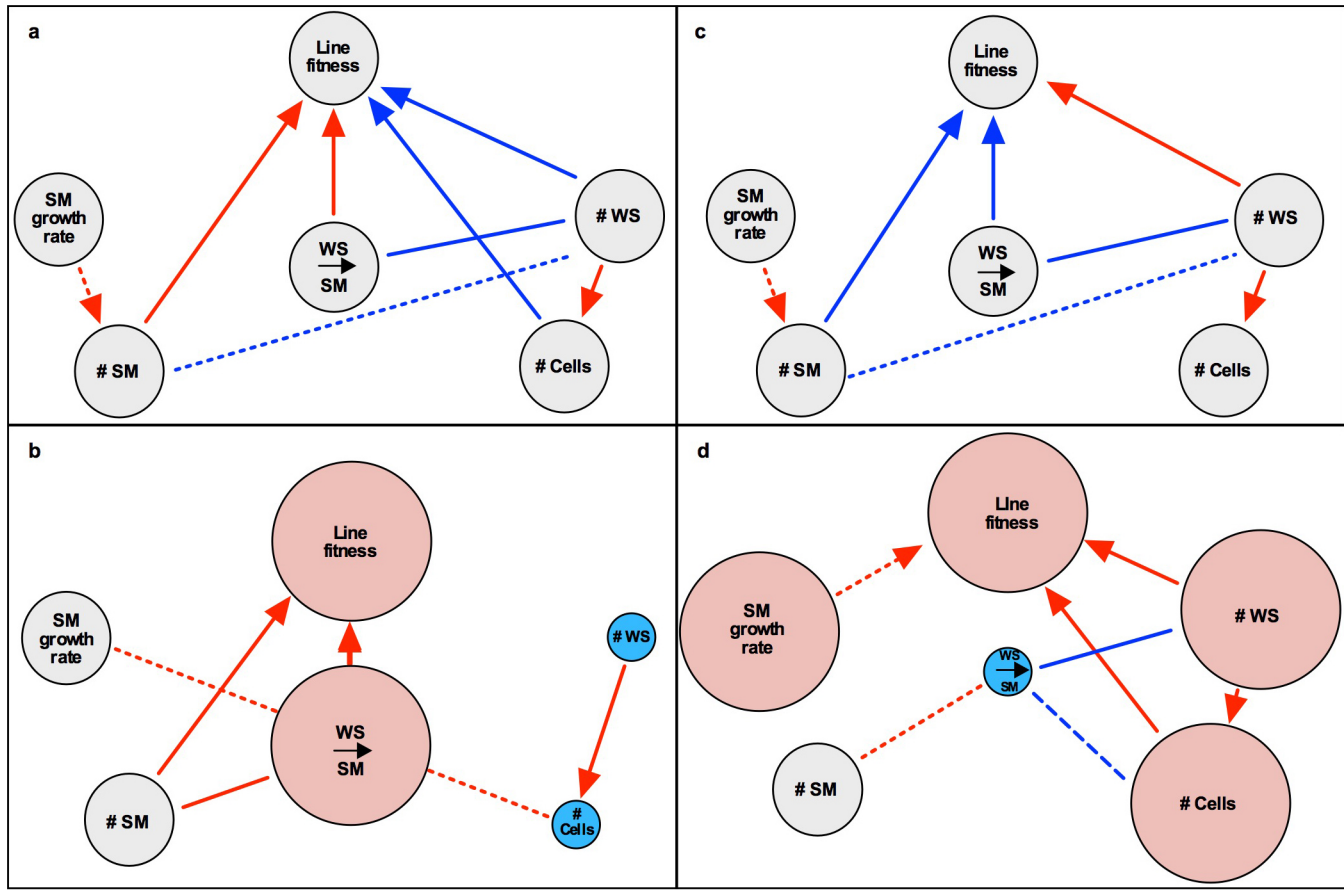
**Extended Data Figure 3 | Life history traits under the CE regime.** a, Cell density of the new type. b, Proportion of the new cell type divided by the total number of cells. Each circle represents the mean of 42–45 lines (that is, three replicates for each of the 15 ancestral and 14 derived lines). Lines that

failed to produce the required type were excluded. Black, derived; grey, ancestral. Error bars are s.e.m., based on  $n \leq 15$ . \* $P < 0.05$ , using analysis of variance (ANOVA) and post hoc contrasts.



**Extended Data Figure 4 | Growth rate of SM.** Growth rate of the ancestral (ANC) and derived SM types from the CE and CP regimes obtained from the representative genotypes (biological and technical replicates; see Methods). (ANC,  $n = 81$ ; CE,  $n = 95$ ; CP  $n = 81$ ). Error bars are s.e.m., based on  $n \leq 15$ .

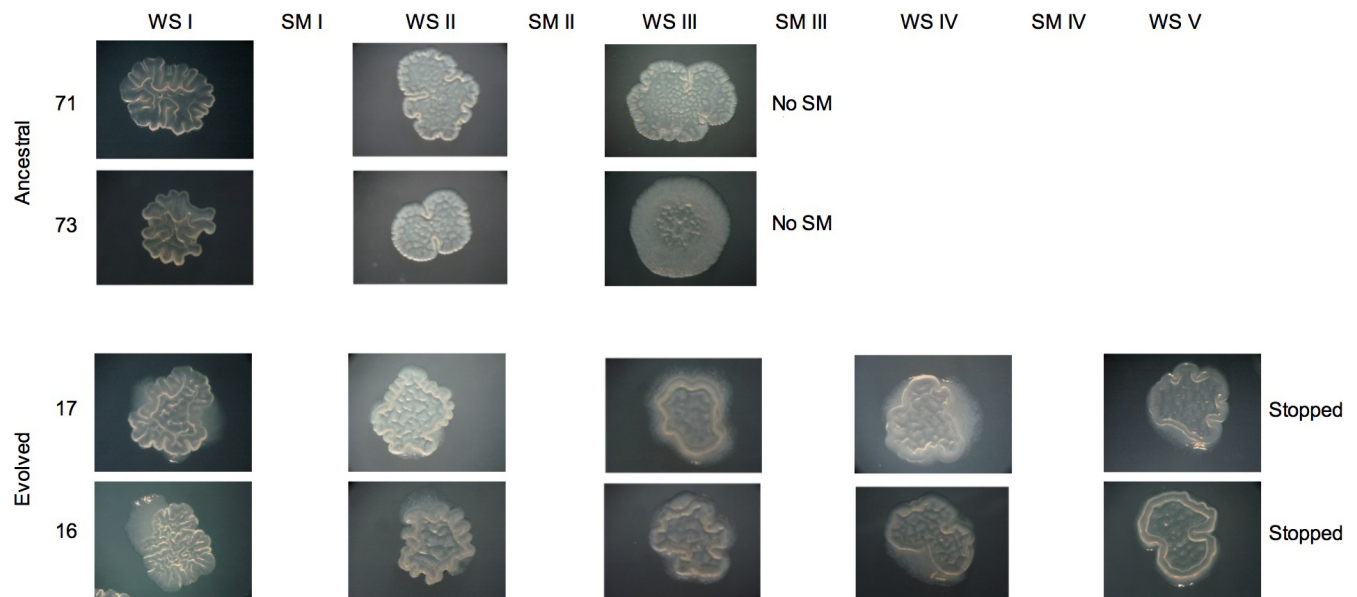




#### Extended Data Figure 5 | Relationship between fitness and associated traits.

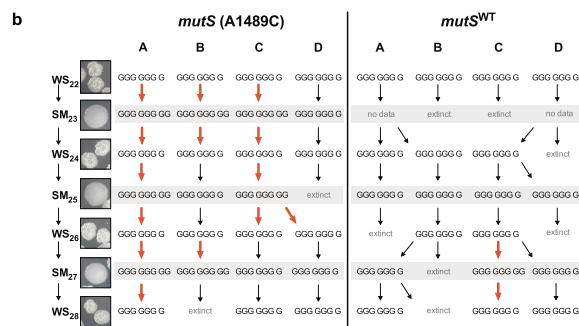
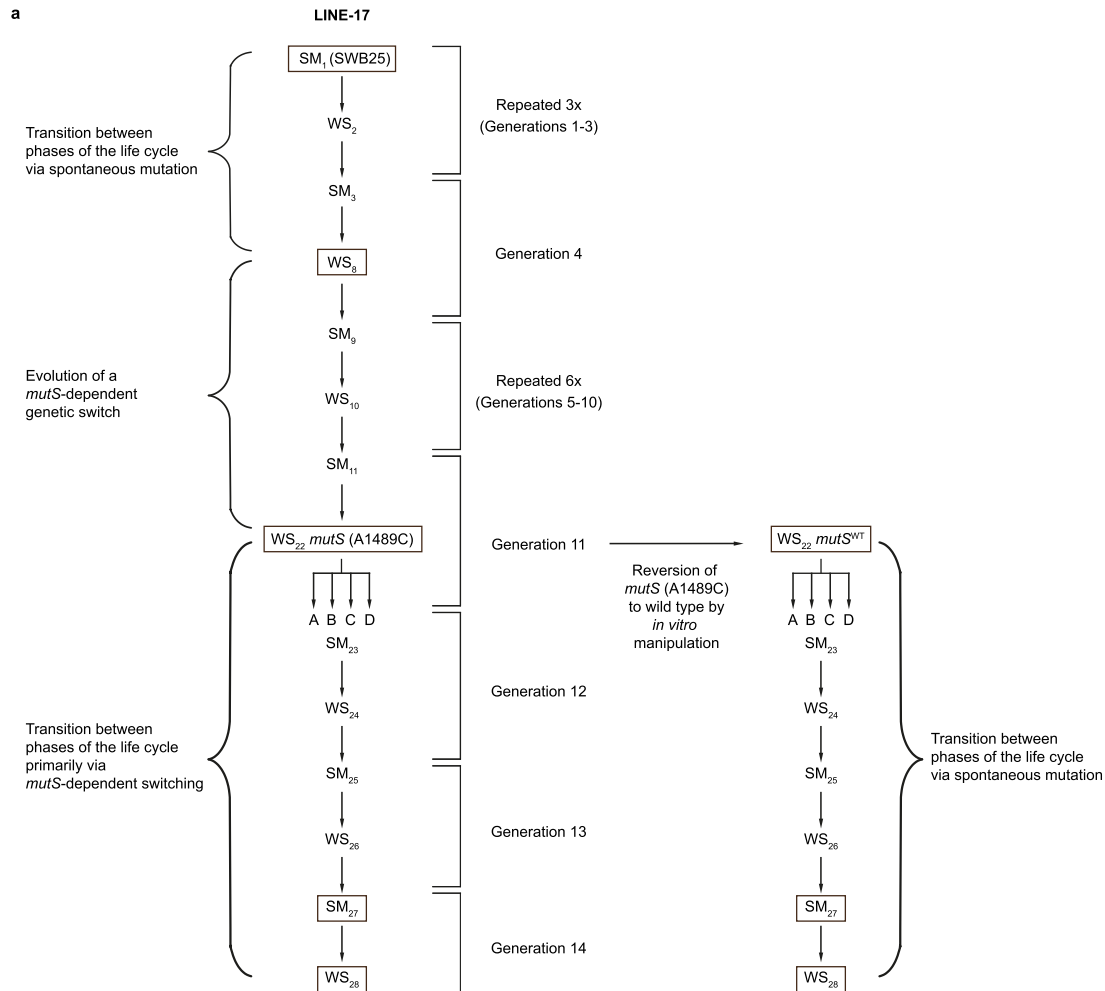
**a–d**, Summary of parameters describing line and cellular properties and the relationship among these parameters in ancestral (**a**) and derived (**b**) CE and ancestral (**c**) and derived (**d**) CP populations. Traits in the evolved populations (**b**, **d**) are depicted relative to their respective ancestral states (**a**, **c**): significant increase (large red circle), significant decrease (small blue circle), and no significant change (grey circle) using a generalized linear model (error structure: binomial; link function: logit) for line fitness and WS→SM with post hoc contrasts; and analysis of variance (ANOVA) for number of cells, number of WS, number of SM, and SM growth rate with post hoc contrasts. WS→SM, proportion of lines producing SM during phase I. Arrows indicate significant regressions and lines indicate significant correlations between traits, dashed lines indicate trends ( $0.05 < P < 0.09$ ). The colour represents the direction of the relationship: red, positive; blue, negative. The significance level is  $P < 0.05$  using Pearson and Spearman rank correlations, and regressions (line level fitness: generalized linear models; cell-level fitness and number of SM per  $\mu\text{l}$  (if present): general linear models). Individual cell properties displayed in **a** and **c** are identical for the ancestral state in both CE and CP regimes, but measures of line fitness are regime-specific, and transform the associations

between parameters. Parameters that relate positively to line fitness in the CE regime negatively affect line fitness in the CP regime, and vice versa (**a** versus **c**). For example, in the CE regime, the number of SM cells and the rate at which WS cells give rise to SM cells positively regresses on line fitness (red arrows, **a**), whereas only the number of WS cells shows a positive regression with line fitness in the CP regime (red arrows, **c**). The relationships between parameters in the ancestral populations predict their evolutionary trajectory in each regime. After 10 generations of line selection the relationships between cell- and line-level parameters significantly altered in both CE and CP regimes (**b** and **d**). Line fitness improved in both regimes, thereby imposing selection on parameters that were linked to line fitness in their respective baselines. In the CE regime enhanced line fitness is explained by a significant increase in the capacity to transition from WS to SM and is not explained by enhanced performance of single cells: the fitness of single cells either remained unaltered or declined. Increased line fitness can be seen as a product of selection at the higher (group) level. In marked contrast is the CP regime where improved line fitness is readily explained by changes in traits that improve the competitive ability of individual cells. Enhanced line fitness in the CP regime can be interpreted as a by-product of selection at the lower (cell) level.



**Extended Data Figure 6 | Cycling through phases.** Ancestral and derived lines differ in their capacity to transition between phases of the life cycle. Three replicate populations of the two ancestral and derived lineages with the highest fitness: 71, 73 and 16, 17, respectively, were founded by the representative WS type and plated to check for SM types. Whereas ancestral WS took 48 h to generate detectable levels of SM, the two derived WS populations contained a mixture of WS and SM colonies, such that even at the time of initial inoculation,

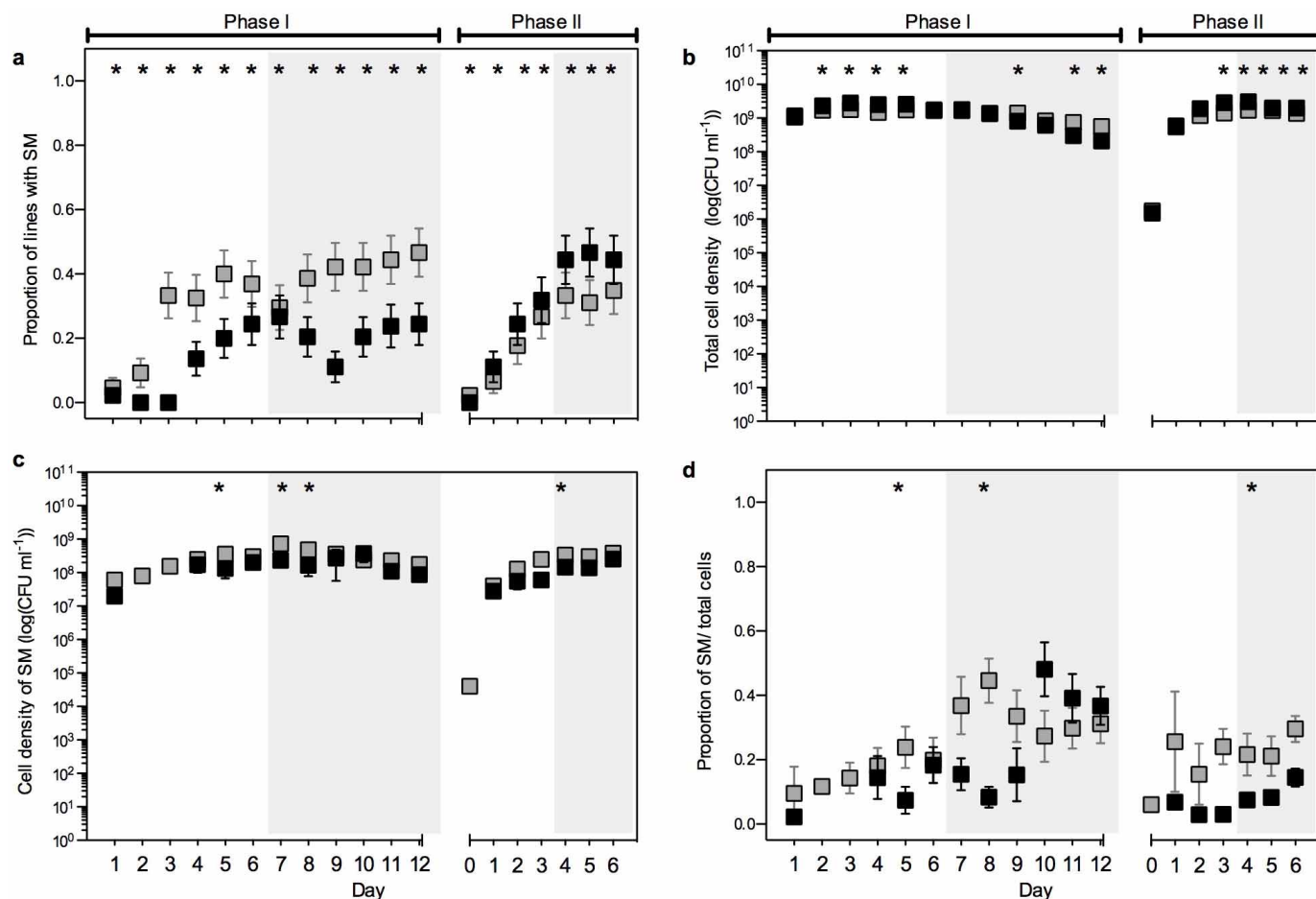
both types were present. SM colonies were then used to found populations that were plated to check for WS types. Ancestral lineages 71 and 73 completed two cycles before extinction through failure to produce SM, whereas derived lines 16 and 17 completed five cycles before termination of the experiment. Colonies were photographed after 48 h of growth. Notable in the evolved lines is the visible presence of zones of SM cells surrounding the central WS colony.



**Extended Data Figure 7 | Mechanism of life cycle transition.** **a**, Outline of the evolutionary history of line 17 from its founding by ancestral SM<sub>1</sub> SBW25 through ten generations of the life cycle, by which time a *mutS* mutation had arisen. At the eleventh generation, the *mutS* (A1489C) mutation in WS<sub>22</sub> was reverted to wild type (*mutS*<sup>WT</sup>) by *in vitro* manipulation. The two WS<sub>22</sub> lineages (with and without the *mutS* mutation) were taken through three additional 'expedited' life-cycle generations, although with a cycling period of 24 h for *mutS* (A1489C) and 48 h for *mutS*<sup>WT</sup>. Genome sequences were acquired from five different time points (SM<sub>1</sub>, WS<sub>8</sub>, WS<sub>22</sub>, SM<sub>27</sub> and WS<sub>28</sub>, indicated by surrounding boxes). For details of mutations see Supplementary Table 1. **b**, The *mutS*-dependent genetic switch. Four independent cultures of the generation-11 WS (WS<sub>22</sub>), with and without the *mutS* (A1489C) mutation, were passed through an additional three 'expedited' generations of the life cycle

(WS<sub>22</sub>–WS<sub>28</sub>). The nucleotide sequence of *wspR* was determined at each stage. Depicted is the tract of guanine residues beginning at nucleotide 742: WspR is active (and the phenotype is WS) when the tract length is seven residues, but inactive (and the phenotype is SM) when the tract length is eight. Red arrows show slippage events resulting in the gain or loss of a single guanine residue. Transitioning between phases is more reliable (fewer extinction events) and more likely to occur via the tract of guanine residues in the presence of *mutS* (A1489C). Lines of *mutS* (A1489C) were passed on a 24-h cycle; *mutS*<sup>WT</sup> lines were passed on a 48-h cycle (on a 24-h cycle *mutS*<sup>WT</sup> lines were extinct before the first generation completed—ancestral lines are incapable of transitioning on a 48-h cycle). Extinction events occurred whenever the extant phase failed to produce the next stage in the life cycle; death of a line allowed birth of an extant lineage (diagonal arrows).





#### Extended Data Figure 8 | Life history traits under the CP regime.

**a**, Proportion of lines producing SM. **b**, Total cell density. **c**, Cell density of the new type. **d**, Proportion of SM cell types divided by the total number of cells. Each square represents the mean of 45 lines (that is, 3 replicates for each of the 15 lines); however, for **c** and **d** lines that failed to produce SM were excluded.

Black, derived; grey, ancestral. Error bars are s.e.m., based on  $n \leq 15$ . \* $P < 0.05$ , using a generalized linear model (error structure: binomial; link function: logit) and post hoc contrasts for **a**; and using analysis of variance (ANOVA) and post hoc contrasts for **b–d**.

Extended Data Table 1 | Differences in life history parameters

a

Parameters	N	df	$F/\chi^2$	P	$R^2$
<b>Line fitness</b>					
			$\chi^2$		
Full model	1392	59	1160.211	<b>&lt;0.0001</b>	
Regime		3	61.429	<b>&lt;0.0001</b>	
Rep. type (Regime)		56	1090.351	<b>&lt;0.0001</b>	
<b>Cell fitness</b>					
			F		
Full model	126	43	4.404	<b>&lt;0.0001</b>	0.698
Regime		2	13.524	<b>&lt;0.0001</b>	
Rep. type (Regime)		41	3.916	<b>&lt;0.0001</b>	
<b>#WS</b>					
			F		
Full model	126	43	5.508	<b>&lt;0.0001</b>	0.743
Regime		2	28.282	<b>&lt;0.0001</b>	
Rep. type (Regime)		41	4.339	<b>&lt;0.0001</b>	
<b>#SM X</b>					
	80		F		
Full model		31	5.148	<b>&lt;0.0001</b>	0.769
Regime		2	2.358	=0.1055	
Rep. type (Regime)		29	5.190	<b>&lt;0.0001</b>	
<b>SM growth rate</b>					
			F		
Full model	104	36	6.450	<b>&lt;0.0001</b>	0.777
Regime		2	32.806	<b>&lt;0.0001</b>	
Rep. type (Regime)		34	5.101	<b>&lt;0.0001</b>	
<b>WS → SM</b>					
			$\chi^2$		
Full model	1557	43	565.727	<b>&lt;0.0001</b>	
Regime		2	289.108	<b>&lt;0.0001</b>	
Rep. type (Regime)		41	467.060	<b>&lt;0.0001</b>	

b

Parameters	N	df	$F/\chi^2$	P	$R^2$
<b>Proportion of lines with new type</b>					
			$\chi^2$		
Full model	1357	64	690.216	<b>&lt;0.0001</b>	
Regime		1	5.442	<b>=0.0197</b>	
Rep. type (Regime)		27	457.619	<b>&lt;0.0001</b>	
Time		18	385.403	<b>&lt;0.0001</b>	
Regime x Time		18	62.994	<b>&lt;0.0001</b>	
<b>Total # cells/<math>\mu</math>l X</b>					
			F		
Full model	1355	46	21.697	<b>&lt;0.0001</b>	0.433
Regime		1	51.521	<b>&lt;0.0001</b>	
Rep. type (Regime)		27	8.144	<b>&lt;0.0001</b>	
Time		18	39.887	<b>&lt;0.0001</b>	
Regime x Time				n.s.	
<b># new type/<math>\mu</math>l X</b>					
			F		
Full model	740	46	8.823	<b>&lt;0.0001</b>	0.369
Regime		1	0.589	=0.4431	
Rep. type (Regime)		27	5.620	<b>&lt;0.0001</b>	
Time		18	13.615	<b>&lt;0.0001</b>	
Regime x Time				n.s.	
<b>Proportion new cell type X</b>					
			F		
Full model	740	46	6.447	<b>&lt;0.0001</b>	0.300
Regime		1	1.701	=0.1926	
Rep. type (Regime)		27	6.648	<b>&lt;0.0001</b>	
Time		18	6.229	<b>&lt;0.0001</b>	
Regime x Time				n.s.	

a. Differences in line fitness, cell fitness and life-history traits between ANC, CE and CP regimes. Results from generalized and general linear models are shown. Values for ancestral lines are the same for all parameters measured during phase I, whereas line fitness differs between ancestral CE and CP regimes. Rep. type, representative WS genotype; #WS, number of WS at day 6 within a mat, #SM X, number of SM at day 6 within a mat (SM = 0 were excluded); X, Box-Cox transformation; WS→SM, proportion of lines producing SM during phase I. b. Differences in life history traits between ancestral and evolved CE regimes over time. Results from generalized and general linear models are shown (all three parameters were Box-Cox transformed). Bold denotes significance at the level of  $P < 0.05$ .

Extended Data Table 2 | Relationship between fitness and life history parameters

a

	SM growth X	#SM/ $\mu$ l X	#WS	WS $\rightarrow$ SM X	Line fitness	Cell fitness
SM growth X	-	$R^2=0.449$ , F=4.893, P=0.0690, N=8	$r=-0.372$ , P=0.2335, N=12	$r=0.222$ , P=0.4882, N=12	$\chi^2=0.711$ , df=1, P=0.3992, N=12	$r=-0.218$ , P=0.4964, N=12
#SM/ $\mu$ l X	$R^2=0.007$ , F=0.087, P=0.773, N=14	-	$r_s=-0.667$ , P=0.0710, N=8	$r_s=0.619$ , P=0.1017, N=8	$\chi^2=26.801$ , df=1, <b>P&lt;0.0001</b> , N=8	$R^2=0.380$ , F <sub>1,8</sub> =3.671, P=0.1038
#WS	$r=0.381$ , P=0.1796, N=14	$r_s=0.007$ , P=0.9822, N=14	-	$r=-0.689$ , <b>P=0.0045</b> , N=15	$\chi^2=11.150$ , df=1, <b>P=0.0008</b> , N=15	<b><math>R^2=0.339</math></b> , F <sub>1,15</sub> =6.673, <b>P=0.0227</b>
WS $\rightarrow$ SM X	$r=0.063$ , P=0.8297, N=14	<b><math>r_s=0.587</math></b> , <b>P=0.0274</b> , <b>N=14</b>	$r=0.114$ , P=0.6987, N=14	-	$\chi^2=22.801$ , df=1, <b>P&lt;0.0001</b> , N=15	$r=-0.378$ , P=0.1645, N=15
Line fitness	$\chi^2=0.061$ , df=1, P=0.8046, N=14	<b><math>\chi^2=9.305</math></b> , <b>df=1</b> , <b>P=0.0023</b> , <b>N=14</b>	$\chi^2=1.008$ , df=1, P=0.3154, N=14	<b><math>\chi^2=12.324</math></b> , <b>df=1</b> , <b>P=0.0004</b> , <b>N=14</b>	-	<b><math>\chi^2=4.246</math></b> , <b>df=1</b> , <b>P=0.0393</b> , <b>N=15</b>
Cell fitness	$r=0.486$ , P=0.0783, N=14	$R^2=0.009$ , F <sub>1,14</sub> =0.113, P=0.7428	<b><math>R^2=0.890</math></b> , <b>F<sub>1,14</sub>=97.359</b> , <b>P&lt;0.0001</b>	$r=0.326$ , P=0.2558, N=14	$\chi^2=2.041$ , df=1, P=0.1531, N=14	-

b

	SM growth X	#SM/ $\mu$ l X	#WS	WS $\rightarrow$ SM X	Line fitness	Cell fitness
SM growth X	-	$R^2=0.449$ , F=4.893, P=0.0690, N=8	$r=-0.372$ , P=0.2335, N=12	$r=0.222$ , P=0.4882, N=12	$\chi^2=0.705$ , df=1, P=0.4010, N=12	$r=-0.218$ , P=0.4964, N=12
#SM/ $\mu$ l X	$R^2=0.011$ , F=0.0656, P=0.8064, N=8	-	$r_s=-0.667$ , P=0.0710, N=8	$r_s=0.619$ , P=0.1017, N=8	$\chi^2=17.568$ , df=1, <b>P&lt;0.0001</b> , N=8	$R^2=0.380$ , F <sub>1,8</sub> =3.671, P=0.1038
#WS	$r=0.215$ , P=0.5265, N=11	$r_s=0.222$ , P=0.5372, N=10	-	$r=-0.689$ , <b>P=0.0045</b> , N=15	$\chi^2=6.153$ , df=1, P=0.0131, N=15	<b><math>R^2=0.339</math></b> , F <sub>1,15</sub> =6.673, <b>P=0.0227</b>
WS $\rightarrow$ SM X	$r=0.221$ , P=0.4129, N=11	$r_s=0.616$ , P=0.0580, N=10	<b><math>r=-0.669</math></b> , <b>P=0.0064</b> , <b>N=15</b>	-	$\chi^2=8.710$ , df=1, <b>P=0.0032</b> , N=15	$r=-0.378$ , P=0.1645, N=15
Line fitness	<b><math>\chi^2=5.567</math></b> , <b>df=1</b> , <b>P=0.0183</b> , <b>N=11</b>	$\chi^2=0.312$ , df=1, P=0.5762, N=10	<b><math>\chi^2=7.552</math></b> , <b>df=1</b> , <b>P=0.0060</b> , <b>N=15</b>	$\chi^2=2.129$ , df=1, P=0.1445, N=15	-	$\chi^2=0.746$ , df=1, P=0.3878, N=15
Cell fitness	$r=0.290$ , P=0.3874, N=11	$R^2=0.074$ , F <sub>1,10</sub> =0.640, P=0.4470	<b><math>R^2=0.887</math></b> , <b>F<sub>1,15</sub>=101.583</b> , <b>P&lt;0.0001</b>	$r=-0.473$ , P=0.0750, N=15	<b><math>\chi^2=5.339</math></b> , <b>df=1</b> , <b>P=0.0208</b> , <b>N=15</b>	-

a, CE regime, and b, CP regime. X denotes parameters that were Box-Cox transformed to meet requirements of normality and equal variance. WS $\rightarrow$ SM, proportion of lines producing SM during phase I. Above the diagonal are tests for the ancestral regimes (grey), below for the evolved regimes (black). Bold denotes significance at the level of  $P < 0.05$ .



# Architecture of mammalian respiratory complex I

Kutti R. Vinothkumar<sup>1\*</sup>, Jiapeng Zhu<sup>2\*</sup> & Judy Hirst<sup>2</sup>

**Complex I (NADH:ubiquinone oxidoreductase) is essential for oxidative phosphorylation in mammalian mitochondria. It couples electron transfer from NADH to ubiquinone with proton translocation across the energy-transducing inner membrane, providing electrons for respiration and driving ATP synthesis. Mammalian complex I contains 44 different nuclear- and mitochondrial-encoded subunits, with a combined mass of 1 MDa. The 14 conserved ‘core’ subunits have been structurally defined in the minimal, bacterial complex, but the structures and arrangement of the 30 ‘supernumerary’ subunits are unknown. Here we describe a 5 Å resolution structure of complex I from *Bos taurus* heart mitochondria, a close relative of the human enzyme, determined by single-particle electron cryo-microscopy. We present the structures of the mammalian core subunits that contain eight iron-sulphur clusters and 60 transmembrane helices, identify 18 supernumerary transmembrane helices, and assign and model 14 supernumerary subunits. Thus, we considerably advance knowledge of the structure of mammalian complex I and the architecture of its supernumerary ensemble around the core domains. Our structure provides insights into the roles of the supernumerary subunits in regulation, assembly and homeostasis, and a basis for understanding the effects of mutations that cause a diverse range of human diseases.**

Mammalian complex I (ref. 1) is one of the largest and most complicated enzymes in the cell. Complex I from *B. taurus* (bovine) heart mitochondria has been characterized extensively as a model for the human enzyme; both enzymes contain 44 different subunits (encoded by both the nuclear and mitochondrial genomes)<sup>2,3</sup> and nine redox cofactors (a flavin mononucleotide and eight iron-sulphur clusters). Fourteen subunits are the core subunits that are conserved in all complex I enzymes; they contain all the mechanistically critical cofactors and structural elements and are sufficient for catalysis. Crystal structures of intact complex I from the thermophilic bacterium *Thermus thermophilus*<sup>4</sup>, and of domains of the prokaryotic enzymes from *T. thermophilus* and *Escherichia coli*<sup>5–7</sup> have provided a wealth of information on the structures of these subunits—but they represent only half the mass of the mammalian enzyme. The cohort of 30 supernumerary subunits particular to the mammalian enzyme<sup>2,3</sup> has been accumulated through evolution. The supernumerary subunits may have alternative functions or be important for assembly, regulation, stability or protection against oxidative stress—their structures and arrangement around the core subunits are not known.

Owing to its size, L-shaped asymmetry, membrane-bound location, and multi-component structure, mammalian complex I has proved difficult to crystallize, and its high-resolution structure has not yet been determined. Crystallographic information on any eukaryotic complex I is currently limited to a medium-resolution map of the enzyme from the yeast *Yarrowia lipolytica*, which has been described, but not modelled<sup>8</sup>. Conversely, the size and shape of complex I make it an attractive target for electron microscopy (EM), and the enzymes from several species have been visualized to display their overall L-shaped structures<sup>9–11</sup>, although at too low a resolution to reveal detailed structural information. A high-resolution structure of the mammalian enzyme is essential for understanding how the 30 supernumerary mammalian subunits are arranged around the core domain, how they determine the properties, assembly and activity of the enzyme, and how mutations in both the core and supernumerary subunits cause human diseases<sup>12</sup>.

## Imaging and reconstruction

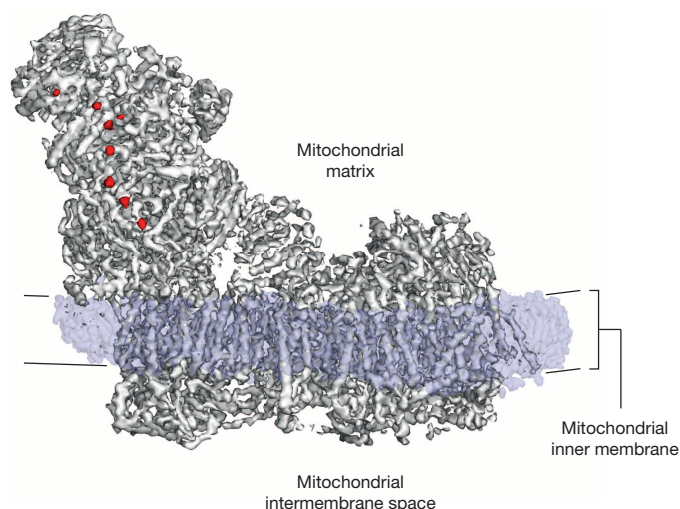
Complex I was purified from *B. taurus* heart mitochondria in detergent<sup>13</sup>, and imaged in vitreous ice on holey-carbon grids with a Falcon direct electron detector (see Methods). The enzyme adopts different orientations on the grid, and reference-free two-dimensional class averages clearly show the characteristic L-shape of the minimal prokaryotic form augmented by extra domains from the supernumerary subunits (Extended Data Fig. 1). Refinement was performed in RELION<sup>14</sup> and movie frames were used to correct for beam-induced movement<sup>15</sup>. Per-frame reconstruction and *B*-factor weighting were followed by three-dimensional classification, resulting in the final map (Fig. 1) obtained from 25,492 particles with an overall resolution of ~5 Å (see Methods and Extended Data Fig. 2). Viewed at a low-density threshold the map is dominated by a disordered detergent-phospholipid belt that encircles the hydrophobic domain and defines the position of the membrane. At intermediate-density threshold, the hydrophilic matrix domain and the extended membrane domain, containing a large number of transmembrane  $\alpha$ -helices (TMHs), are observed. The highest-density peaks in the map reveal the eight iron-sulphur (FeS) clusters that, as in the *T. thermophilus*<sup>4,7</sup> and *Y. lipolytica*<sup>8</sup> enzymes, form a chain through the hydrophilic domain.

## Structures of the core subunits

The 14 conserved core subunits of complex I (refs 1, 4) catalyse the energy transducing reactions: NADH oxidation, ubiquinone reduction and proton translocation (Extended Data Table 1 summarizes their nomenclature). The seven nuclear-encoded hydrophilic core subunits harbour a flavin mononucleotide to oxidize NADH, FeS clusters for inter-substrate electron transfer, and the ubiquinone-binding site. The seven mitochondrial-encoded membrane core subunits contain four antiporter-like domains for proton translocation. The structures of the mammalian core subunits (Fig. 2) were fitted to the density map (see Methods) using the structure of *T. thermophilus* complex I (ref. 4), secondary structure analyses and sequence alignments, and using structural features and

<sup>1</sup>MRC Laboratory of Molecular Biology, Francis Crick Avenue, Cambridge CB2 0QH, UK. <sup>2</sup>MRC Mitochondrial Biology Unit, Wellcome Trust/MRC Building, Hills Road, Cambridge CB2 0XY, UK.

\*These authors contributed equally to this work.

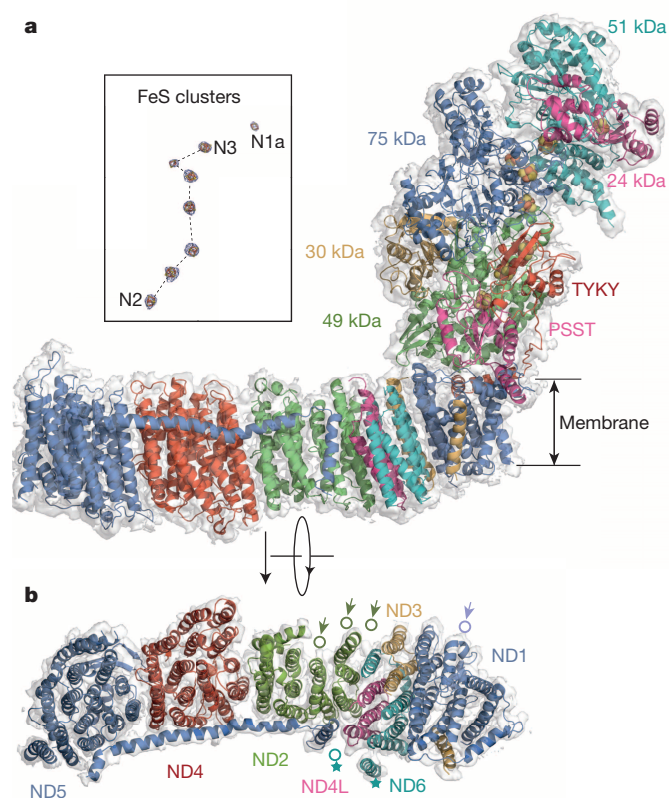


**Figure 1 | Overall map for complex I from *B. taurus* heart mitochondria determined by single-particle cryo-EM.** Three distinct features of the complex are revealed by overlaying maps at different density thresholds. The map at the highest threshold (red) reveals the FeS clusters. The map at medium threshold (grey) reveals the overall architecture of the protein and the 78 TMHs in the membrane domain. The detergent-phospholipid belt observed as a dominant feature at low density threshold (translucent blue) represents the density that remains around the membrane domain after cutting out the final model of the protein, and denotes the position of the complex in the membrane. It is ~30 Å thick, and 3–4 Å thinner at the proximal end of the complex (left) than at the distal end (right).

densities from aromatic side chains (Extended Data Fig. 3). Except for the FeS-cluster ligands they have been modelled as polyalanine chains, with the residue numbering optimized to enable individual residues to be located (Extended Data Table 2). It is not possible to attribute density to any bound ubiquinone species in the present map.

A comparison of the bacterial<sup>4</sup> and mammalian core enzymes reveals that the mammalian membrane domain is more strongly curved 'out' of the membrane plane (Extended Data Fig. 4). However, within each individual subunit the 60 TMHs of the mammalian core subunits match their *T. thermophilus* counterparts closely (Extended Data Fig. 5)—only the position of TMH4 in subunit ND6 is different, and the extra carboxy-terminal TMH particular to *T. thermophilus* ND1 is absent from *B. taurus* (Fig. 2 and Extended Data Fig. 5). No notable density is observed in place of the three amino-terminal TMHs (present in *T. thermophilus* and *Y. lipolytica*) that have been lost through evolution of mammalian ND2 (ref. 16), so they have not been substituted structurally by other subunits. Importantly, catalytically relevant features identified in the antiporter-like subunits of the bacterial complex<sup>5,6</sup> are conserved. They include the loops in the six broken TMHs in ND2, ND4 and ND5 (see Extended Data Fig. 3 for examples) that may constitute part of the proton-translocation mechanism, and the long transverse helix in ND5, a proposed coupling element.

Of the seven hydrophilic core subunits (Fig. 2), the structures of the *B. taurus* 51 kDa subunit (human homologue NDUFV1), 49 kDa (NDUFS2), 24 kDa (NDUFV2), PSST (NDUFS7) and TYKY (NDUFS8) subunits, and the small domain of the 75 kDa subunit (NDUFS1) are closely conserved from their *T. thermophilus* homologues<sup>4,7</sup> (Extended Data Fig. 5), with marked variation only in the length and extent of some of their N and C termini. Consequently, the arrangements of the FeS cluster chains are also very similar (Extended Data Table 3), except that, owing to rotation of the 51 and 24 kDa subunits, the superimposed chains diverge with increasing distance from the membrane (Extended Data Fig. 4). The sequence and structural conservation of the large domain of the 75 kDa subunit, which contains an extra, catalytically redundant cluster in *T. thermophilus*<sup>7</sup> and the 30 kDa subunit (NDUFS3), are lower (Extended Data Table 2). As neither of them have any known catalytic



**Figure 2 | Structures of the core subunits of mammalian complex I.**

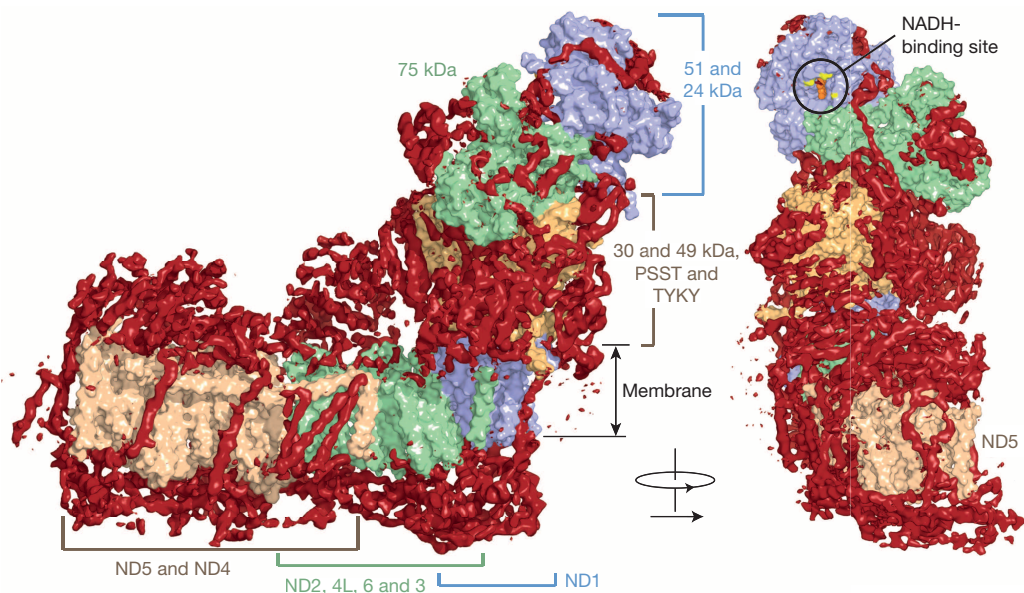
**a**, Structural models of the 14 mammalian core subunits (cartoon representation) and their density (transparent surface); the subunits are coloured individually and labelled with text in the same colours. The chain of FeS clusters is shown modelled to the highest density peaks (blue mesh) in the inset. **b**, The seven membrane-bound mammalian core subunits, viewed from the matrix. Arrows indicate the positions of the four TMHs in *T. thermophilus* that are not present in *B. taurus*: three N-terminal TMHs in ND2 and one C-terminal TMH in ND1. The position of TMH4 in ND6 is different in *B. taurus* and *T. thermophilus* (marked with stars). For a detailed comparison of the *B. taurus* and *T. thermophilus* structures see Extended Data Figs 4 and 5.

role, we conclude that the catalytically critical subunits and cofactors are closely conserved in the mitochondrial and bacterial enzymes, supporting their common mechanism of catalysis.

## The supernumerary ensemble

Once the core subunits had been modelled, the map revealed that additional densities form an open cage around the core (Fig. 3). These densities are attributed to the supernumerary subunits, and they are arranged predominantly around the membrane domain and lower hydrophilic domain, where they may help to protect FeS-containing PSST and TYKY from oxidative damage. Conversely, the area around the NADH-binding site, where complex I produces superoxide<sup>17</sup>, is bare (Fig. 3), so supernumerary subunits do not shield it from O<sub>2</sub> to minimize superoxide production. The NADH dehydrogenase domain is added at the end of the complex I assembly pathway<sup>18</sup>, and the local paucity of supernumerary subunits may facilitate both its integration and its replacement (while retaining the rest of the protein) to mitigate the effects of oxidative damage<sup>19</sup>. Two large supernumerary domains, one capping ND5 and part of ND4, the other capping ND2, are observed on the matrix surface of the membrane domain. Facing the intermembrane space, as noted in *Y. lipolytica*<sup>8</sup>, the supernumerary subunits form a layer of protein that may have a role similar to that of the stabilizing  $\beta$ -hairpin-helix structures observed in the prokaryotic enzyme<sup>5</sup>. Eighteen supernumerary TMHs are distributed around the core membrane domain





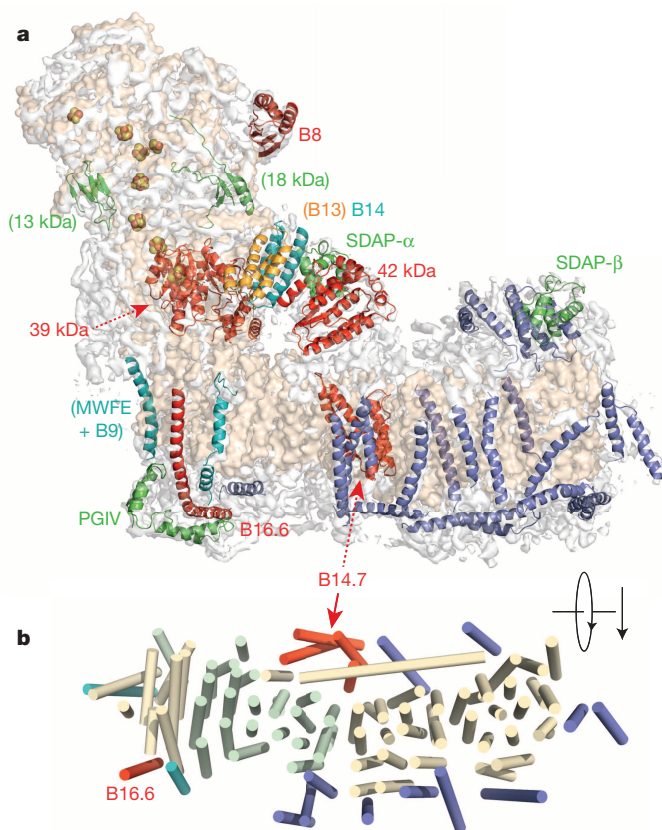
**Figure 3 | Architecture of mammalian complex I showing the densities of the supernumerary subunits enclosing the core domain.** The models for the core subunits are in light colours (as labelled) in surface representation, and density attributed to the supernumerary subunits, forming a cage around the core subunits, is in dark red. The supernumerary subunits are concentrated on each side of the membrane domain, and around the lower section of the hydrophilic domain. The NADH-binding site in the 51 kDa subunit is indicated, with the predicted positions for the flavin isoalloxazine (orange spheres) and three conserved phenylalanines at the entry to the site (yellow); the vicinity of this site is devoid of supernumerary subunit density.

(Figs 3 and 4), consistent with the predictions of sequence analyses for 14 to 18 TMHs from these subunits (Extended Data Table 4). In total, therefore, we observe 78 TMHs in the mammalian enzyme. Two TMHs are on the outside of the ND5 transverse helix, appearing to strap it to the core domain, and four more are positioned close to the end of it, appearing as a restraint for its lateral movement (Fig. 3). These observations raise the question of whether large-scale piston-like motions of this helix during catalysis, as postulated from the *T. thermophilus* structure<sup>6</sup>, are feasible.

### Assignment of 14 supernumerary subunits

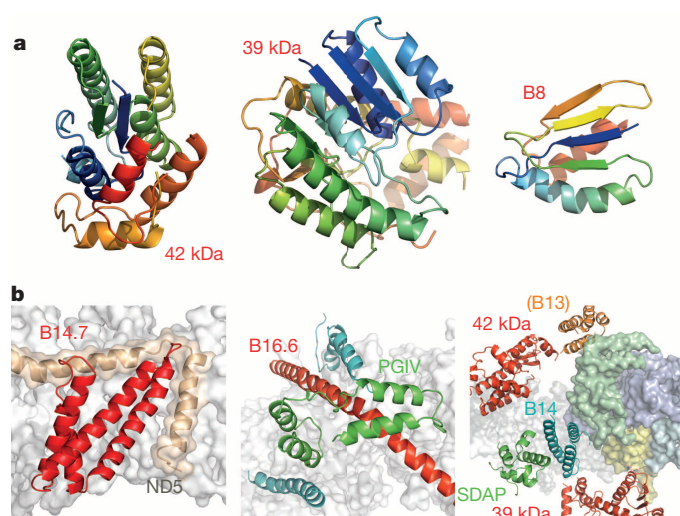
To identify and assign individual supernumerary subunits to the map for mammalian complex I (Extended Data Table 4 summarizes their nomenclature) we used biochemical, sequence and structural information. Homology models for six of the hydrophilic supernumerary subunits were created using known structures (Extended Data Tables 4 and 5). Human B8 (NDUFA2) adopts a thioredoxin fold<sup>20</sup> and its structure (Fig. 5) was located at the tip of the large domain of the 75 kDa subunit (Fig. 4), so (contrary to current models<sup>18</sup>) B8 is likely to be assembled into complex I after (or with) the 75 kDa subunit. B8 is extensively degraded in brain mitochondria from patients with Parkinson's disease<sup>21</sup>, and, along with other NADH dehydrogenase domain subunits, it is rapidly exchanged under steady-state conditions<sup>19</sup>. Therefore, it may help to protect the core enzyme against oxidative damage. Similarly, regions of density consistent with two subunits important for complex I assembly<sup>18</sup>, the 18 kDa (NDUFS4) and 13 kDa (NDUFS6) subunits (Extended Data Tables 4 and 5), were located (Fig. 4). However, they are small proteins with no predicted dominant secondary structure and it cannot be excluded that other supernumerary subunits have similar structures. In the current map, the 18 kDa subunit has been modelled into a density in a cleft between the 75 kDa subunit and the 49 kDa, 30 kDa and TYKY subunits; the density attributed to the 13 kDa subunit suggests that it interacts with the 75 kDa, 49 kDa and TYKY subunits (Fig. 4). These locations may explain why clinically identified mutations in the 18 kDa and 13 kDa subunits lead to accumulation of late-stage interrupted-assembly intermediates lacking the NADH-dehydrogenase module<sup>22,23</sup>.

The 42 kDa subunit (NDUFA10), a member of the nucleoside kinase family<sup>24</sup>, was easily located as the density on top of ND2, on the matrix side of the membrane (Figs 4, 5 and Extended Data Tables 4, 5). Its location is neatly confirmed by its absence from the density map of *Y. lipolytica* complex I (ref. 8), which lacks this mammalian-specific subunit. Phosphorylation of a serine in the 42 kDa subunit by a PINK1-dependent mechanism has been proposed to be required for complex I



**Figure 4 | Structural assignments of supernumerary subunits in mammalian complex I.** **a**, A semi-transparent surface for the density map for mammalian complex I is shown in pale grey, with the surface from the core subunits in wheat. Structural models for the supernumerary subunits are shown in colour and labelled accordingly (dashed lines indicate subunits on the back of the structure). Subunits labelled with brackets are those with less certain assignments, and structural elements, which cannot be assigned confidently in the current map, are in blue. **b**, Arrangement of TMHs, viewed from the matrix. The core subunits are in light colours (wheat for ND1, ND4 and ND5, green for ND2, ND3, ND4L and ND6). The supernumerary subunits are coloured as in **a**.





**Figure 5 | Structural models for supernumerary subunits in mammalian complex I.** **a**, Models for three supernumerary subunits in cartoon representation, coloured from blue to red (N to C termini). **b**, Structural models and relationships of supernumerary subunits to the core structure. B14.7 is located at the end of the transverse helix, next to ND5–TMH16. The density assigned to PGIV forms an L-shaped ‘clip’ over B16.6, which bends around the heel at ND1. Finally, the supernumerary subunits around the lower section of the hydrophilic domain are viewed in cartoon representation from the matrix. The core membrane subunits (white) and four core hydrophilic subunits, the 49 kDa (blue), 30 kDa (green), PSST (yellow) and TYKY (cyan) subunits, are shown in surface representation.

activity<sup>25</sup>, implying both its regulatory role and a molecular link between PINK1 dysfunction and complex I activity in Parkinson’s disease. Further elucidation of this regulatory pathway must now be reconciled with the matrix location of the 42 kDa subunit. The 39 kDa subunit (NDUFA9), a member of the nucleotide-binding short-chain dehydrogenase/reductase family<sup>26</sup> (Extended Data Tables 4 and 5), was readily located adjacent to PSST (Figs 4 and 5), and observed to contain a density consistent with a bound dinucleotide (Extended Data Fig. 3)<sup>27</sup>. Furthermore, it partially encloses the long ND3 loop (resolved only in *T. thermophilus*<sup>4</sup>) that is critical for coupling electron and proton transfer, and which, in a conformational transition now known to also involve the 39 kDa subunit<sup>28</sup>, switches the enzyme into a ‘deactive’ state during ischaemia. Notably, these proposed regulatory elements are all located close to the junction between the hydrophilic and membrane domains where the energy from the redox reaction is used to initiate proton translocation<sup>1</sup>.

Two regions of density corresponding to the structure of the SDAP subunit (NDUFAB1), which is identical to the acyl carrier protein in the mitochondrial matrix<sup>29,30</sup>, were identified in the mammalian enzyme (Fig. 4). This result is supported by the presence of SDAP in both subcomplex I $\alpha$  (which contains the hydrophilic domain) and subcomplex I $\beta$  (the distal portion of the membrane domain) of *B. taurus* complex I (ref. 3), and with the presence of two SDAP homologues in *Y. lipolytica* complex I (ref. 31). One SDAP is located at the distal end of the enzyme, above ND5, the other in a peripheral region of the hydrophilic domain, in a subdomain that interacts with the 49 kDa and 30 kDa core subunits through a three-helix bundle (Figs 4 and 5). From a recent study in *Y. lipolytica*<sup>32</sup> these helices are assigned to subunit B14 (NDUFA6), a protein with an LYR motif that, when deleted in *Y. lipolytica*, results in loss of catalytic activity. Notably, subunit B22 (NDUFB9) also contains an LYR motif and it is in subcomplex I $\beta$ , so it is possible that the distal SDAP molecule interacts with it in a similar fashion. Finally, subunits B13 (NDUFA5) and B14 have similar predicted secondary structures so we ascribe the second three-helix bundle observed on the side of the 30 kDa subunit, adjacent to the 42 kDa subunit, to B13 (Figs 4, 5).

Continuous density links the four supernumerary TMHs at the end of the transverse helix. Only one subunit, B14.7 (NDUFA11), is predicted to contain more than two TMHs (Extended Data Table 4); secondary structure analyses predict that the first two are unusually long (~30 residues), so they probably correspond to the two highly tilted TMHs. Therefore, these four TMHs are assigned to subunit B14.7 (Figs 4, 5 and Extended Data Table 5), a protein that is important for the assembly and/or stability of the membrane domain<sup>33</sup>. A second cluster of three TMHs (opposite B14.7) may include two TMHs from B14.5b (NDUFC2), but the connectivity between them is ambiguous and a clear assignment cannot be made. The 11 remaining TMHs are spread around the membrane domain (Fig. 4). Three TMH-containing subunits, B16.6 (NDUFA13), MWFE (NDUFA1) and B9 (NDUFA3), remain associated with the hydrophilic domain in subcomplex I $\alpha$  after fractionation of *B. taurus* complex I with zwitterionic detergents<sup>3</sup>, and they are missing from *Y. lipolytica* subcomplex I $\delta$ , which lacks core subunits ND1, 2, 3 and 4L<sup>34</sup>. Therefore, they are assigned to the three TMH densities next to ND1 (Extended Data Table 4 and 5). Sequence analyses predict a single 67-residue helix in subunit B16.6, with the first 20 residues forming a TMH. Correspondingly, one of the three densities is very long and modelled as a single 63-residue helix that interacts with the N terminus of TYKY on the matrix side, spans the membrane, then bends into the intermembrane space and is anchored under the ‘heel’ (Figs 4 and 5). Therefore, this density is assigned to B16.6, a protein identical to the cell death regulatory gene product GRIM19 (ref. 35). It is currently not possible to confidently deduce assignments for the TMH-containing subunits of subcomplex I $\beta$ .

The PGIV (NDUFA8), 15 kDa (NDUFS5) and B18 (NDUFB7) subunits contain twin CX<sub>9</sub>C motifs that form two intramolecular disulphides within ‘CHCH’ domains<sup>36</sup>, and they have been assigned to the intermembrane surface of complex I (ref. 37). A double L-shaped density, resembling two CHCH domains at right angles, is clearly visible on the heel, clamping B16.6 (Figs 4 and 5) onto the core. PGIV, a subunit present in subcomplex I $\alpha$  (Extended Data Table 4), contains two CHCH domains, so it is assigned to the L-shaped density feature (Figs 4 and 5), consistent with the position of an antibody label to its homologous subunit in *Y. lipolytica*<sup>34</sup>. Our structure thus reveals the architecture of the ‘400 kDa’ assembly intermediate of human complex I (refs 18, 33) that contains the core hydrophilic subunits 49 kDa, 30 kDa, PSST and TYKY, core membrane subunit ND1, and the supernumerary subunits PGIV, B9, B16.6 and B13.

## Conclusions and perspectives

We have described a 5 Å resolution cryo-EM density map for mammalian complex I, and used it to produce structural models for the 14 core subunits that are conserved in all complex I enzymes, plus 14 of the supernumerary subunits of the mammalian enzyme. The core subunits comprise the catalytically active centre of the enzyme, and (as expected) they closely resemble their counterparts described by the atomic-resolution structure of bacterial complex I (ref. 4). The 14 supernumerary subunits assigned include two copies of subunit SDAP, bringing the total number of subunits in the mammalian complex up to 45. We have used our structural models for the supernumerary subunits to support and discuss their roles in assembly, homeostasis and regulation. Higher-resolution maps are required for assignment of the remaining 17 supernumerary subunits.

Recent developments in direct electron detectors, microscopy and image processing algorithms have enabled high-resolution structures of biological macromolecules to be determined by single-particle cryo-EM at resolutions that have previously only been routinely possible with X-ray crystallography<sup>38–40</sup>. Thus, we believe that it will be possible to extend our current study to produce a high-resolution structure for complex I in the near future, to allow us to identify and model all the supernumerary subunits, and to characterize the structural changes that occur during catalysis, a crucial step in defining the mechanism of electron-coupled proton translocation.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 14 April; accepted 17 July 2014.**

**Published online 7 September 2014.**

- Hirst, J. Mitochondrial complex I. *Annu. Rev. Biochem.* **82**, 551–575 (2013).
- Carroll, J., Fearnley, I. M., Shannon, R. J., Hirst, J. & Walker, J. E. Analysis of the subunit composition of complex I from bovine heart mitochondria. *Mol. Cell. Proteomics* **2**, 117–126 (2003).
- Hirst, J., Carroll, J., Fearnley, I. M., Shannon, R. J. & Walker, J. E. The nuclear encoded subunits of complex I from bovine heart mitochondria. *Biochim. Biophys. Acta* **1604**, 135–150 (2003).
- Baradaran, R., Berrisford, J. M., Minhas, G. S. & Sazanov, L. A. Crystal structure of the entire respiratory complex I. *Nature* **494**, 443–448 (2013).
- Efremov, R. G. & Sazanov, L. A. Structure of the membrane domain of respiratory complex I. *Nature* **476**, 414–420 (2011).
- Efremov, R. G., Baradaran, R. & Sazanov, L. A. The architecture of respiratory complex I. *Nature* **465**, 441–445 (2010).
- Sazanov, L. A. & Hinchliffe, P. Structure of the hydrophilic domain of respiratory complex I from *Thermus thermophilus*. *Science* **311**, 1430–1436 (2006).
- Hunte, C., Zickermann, V. & Brandt, U. Functional modules and structural basis of conformational coupling in mitochondrial complex I. *Science* **329**, 448–451 (2010).
- Leonard, K., Haiker, H. & Weiss, H. Three-dimensional structure of NADH:ubiquinone reductase (complex I) from *Neurospora mitochondria* determined by electron microscopy of membrane crystals. *J. Mol. Biol.* **194**, 277–286 (1987).
- Grigorieff, N. Three-dimensional structure of bovine NADH:ubiquinone oxidoreductase (complex I) at 2.2 Å in ice. *J. Mol. Biol.* **277**, 1033–1046 (1998).
- Clason, T. *et al.* The structure of eukaryotic and prokaryotic complex I. *J. Struct. Biol.* **169**, 81–88 (2010).
- Fassone, E. & Rahman, S. Complex I deficiency: clinical features, biochemistry and molecular genetics. *J. Med. Genet.* **49**, 578–590 (2012).
- Sharpley, M. S., Shannon, R. J., Draghi, F. & Hirst, J. Interactions between phospholipids and NADH:ubiquinone oxidoreductase (complex I) from bovine mitochondria. *Biochemistry* **45**, 241–248 (2006).
- Scheres, S. H. W. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **180**, 519–530 (2012).
- Bai, X.-C., Fernandez, I. S., McMullan, G. & Scheres, S. H. W. Ribosome structures to near-atomic resolution from thirty thousand cryo-EM particles. *eLife* **2**, e00461 (2013).
- Birrell, J. A. & Hirst, J. Truncation of subunit ND2 disrupts the threefold symmetry of the antiporter-like subunits in complex I from higher metazoans. *FEBS Lett.* **584**, 4247–4252 (2010).
- Kusssmaul, L. & Hirst, J. The mechanism of superoxide production by NADH:ubiquinone oxidoreductase (complex I) from bovine heart mitochondria. *Proc. Natl Acad. Sci. USA* **103**, 7607–7612 (2006).
- Mimaki, M., Wang, X., McKenzie, M., Thorburn, D. R. & Ryan, M. T. Understanding mitochondrial complex I assembly in health and disease. *Biochim. Biophys. Acta* **1817**, 851–862 (2012).
- Dieteren, C. E. J. *et al.* Subunit-specific incorporation efficiency and kinetics in mitochondrial complex I homeostasis. *J. Biol. Chem.* **287**, 41851–41860 (2012).
- Brockmann, C. *et al.* The oxidised subunit B8 from human complex I adopts a thioredoxin fold. *Structure* **12**, 1645–1654 (2004).
- Keeney, P. M., Xie, J., Capaldi, R. A. & Bennett, J. P. Parkinson's disease brain mitochondrial complex I has oxidatively damaged subunits and is functionally impaired and misassembled. *J. Neurosci.* **26**, 5256–5264 (2006).
- Leshinsky-Silver, E. *et al.* NDUFS4 mutations cause Leigh syndrome with predominant brainstem involvement. *Mol. Genet. Metab.* **97**, 185–189 (2009).
- Kirby, D. M. *et al.* NDUFS6 mutations are a novel cause of lethal neonatal mitochondrial complex I deficiency. *J. Clin. Invest.* **114**, 837–845 (2004).
- Sharpley, M. S. *Studies of the Catalytic Activity of NADH:Ubiquinone Oxidoreductase (Complex I) from Bovine Mitochondria*. PhD thesis, Cambridge Univ. (2005).
- Morais, V. A. *et al.* PINK1 loss of function mutations affect mitochondrial complex I activity via NdufA10 ubiquinone uncoupling. *Science* **344**, 203–207 (2014).
- Fearnley, I. M. & Walker, J. E. Conservation of sequences of subunits of mitochondrial complex I and their relationships with other proteins. *Biochim. Biophys. Acta* **1140**, 105–134 (1992).
- Abdrakhmanova, A., Zwicker, K., Kersch, S., Zickermann, V. & Brandt, U. Tight binding of NADPH to the 39-kDa subunit of complex I is not required for catalytic activity but stabilizes the multiprotein complex. *Biochim. Biophys. Acta* **1757**, 1676–1682 (2006).
- Babot, M. *et al.* ND3, ND1 and 39 kDa subunits are more exposed in the de-active form of bovine mitochondrial complex I. *Biochim. Biophys. Acta* **1837**, 929–939 (2014).
- Runswick, M. J., Fearnley, I. M., Skehel, J. M. & Walker, J. E. Presence of an acyl carrier protein in NADH:ubiquinone oxidoreductase from bovine heart mitochondria. *FEBS Lett.* **286**, 121–124 (1991).
- Cronan, J. E., Fearnley, I. M. & Walker, J. E. Mammalian mitochondria contain a soluble acyl carrier protein. *FEBS Lett.* **579**, 4892–4896 (2005).
- Dobrynin, K. *et al.* Characterization of two different acyl carrier proteins in complex I from *Yarrowia lipolytica*. *Biochim. Biophys. Acta* **1797**, 152–159 (2010).
- Angerer, H. *et al.* The LYR protein subunit NB4M/NDUFA6 of mitochondrial complex I anchors an acyl carrier protein and is essential for catalytic activity. *Proc. Natl Acad. Sci. USA* **111**, 5207–5212 (2014).
- Andrews, B., Carroll, J., Ding, S., Fearnley, I. M. & Walker, J. E. Assembly factors for the membrane arm of human complex I. *Proc. Natl Acad. Sci. USA* **110**, 18934–18939 (2013).
- Angerer, H. *et al.* A scaffold of accessory subunits links the peripheral arm and the distal proton-pumping module of mitochondrial complex I. *Biochem. J.* **437**, 279–288 (2011).
- Fearnley, I. M. *et al.* GRIM-19, a cell death regulatory gene product, is a subunit of bovine mitochondrial NADH:ubiquinone oxidoreductase (complex I). *J. Biol. Chem.* **276**, 38345–38348 (2001).
- Banci, L. *et al.* Structural characterization of CHCHD5 and CHCHD7: two atypical human twin CX<sub>9</sub>C proteins. *J. Struct. Biol.* **180**, 190–200 (2012).
- Szklarczyk, R. *et al.* NDUFB7 and NDUFA8 are located at the intermembrane surface of complex I. *FEBS Lett.* **585**, 737–743 (2011).
- Liao, M., Cao, E., Julius, D. & Cheng, Y. Structure of the TRPV1 ion channel determined by electron cryo-microscopy. *Nature* **504**, 107–112 (2013).
- Amunts, A. *et al.* Structure of the yeast mitochondrial large ribosomal subunit. *Science* **343**, 1485–1489 (2014).
- Allegretti, M., Mills, D. J., McMullan, G., Kühlbrandt, W. & Vonck, J. Atomic model of the F<sub>420</sub>-reducing [NiFe] hydrogenase by electron cryo-microscopy using a direct electron detector. *eLife* **3**, e01963 (2014).

**Acknowledgements** We thank R. Henderson, S. H. W. Scheres, G. McMullan, G. Murshudov, P. Emsley and J. E. Walker for helpful advice, the FEI fellows for educating us on use of the Titan Krios, J. Grimmett and T. Darling for computational help, and S. Chen and C. Savva for EM help. This work was supported by the Medical Research Council, grant numbers U105184322 (K.R.V., in R. Henderson's group) and U105663141 (J.H.).

**Author Contributions** K.R.V. carried out EM experiments and analysis; J.Z. prepared protein; K.R.V., J.Z. and J.H. modelled and analysed data; J.H. designed the project; K.R.V., J.Z. and J.H. wrote the paper.

**Author Information** The EM map of complex I has been deposited in the Electron Microscopy Data Bank under accession number EMD-2676, and the associated model has been deposited in the Protein Data Bank under accession number 4UQ8. Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to K.R.V. ([vkumar@mrc-lmb.cam.ac.uk](mailto:vkumar@mrc-lmb.cam.ac.uk)) or J.H. ([jh@mrc-mbu.cam.ac.uk](mailto:jh@mrc-mbu.cam.ac.uk)).

## METHODS

**Protein preparation.** Complex I was purified from *B. taurus* heart mitochondrial membranes by solubilization and anion exchange chromatography in *n*-dodecyl- $\beta$ -D-maltoside (DDM), and size-exclusion chromatography in DDM or 7-cyclohexyl-1-heptyl- $\beta$ -D-maltoside (Cymal 7) as described previously<sup>13</sup>.

**Cryo-EM specimen preparation and imaging.** Aliquots of complex I (3  $\mu$ l, 3–4.5 mg ml<sup>-1</sup>) were applied to glow-discharged holey-carbon Quantifoil R 0.6/1 grids, blotted for 15–18 s, then plunge-frozen in liquid ethane using an environmental plunge-freeze apparatus<sup>41</sup>. The grids were transferred into cartridges, loaded into an FEI Titan Krios electron microscope, and images were recorded at 2–5  $\mu$ m under-focus on a Falcon II CMOS (complementary metal oxide semiconductor) direct electron detector at 300 keV at  $\times 81,495$  magnification (nominally  $\times 47,000$ ), with the specimen temperature at  $-186^\circ\text{C}$  using the EPU software (Extended Data Fig. 1). The detector pixel size of 14  $\mu$ m corresponds to a sampling density of  $\sim 1.72$  Å pixel<sup>-1</sup>. Each image was exposed for 4 s (total dose  $\sim 64$  e Å<sup>-2</sup>) and 72 frames were captured as previously described<sup>15</sup>. For tilt-pair analysis, the same area was imaged twice, at  $0^\circ$  for 0.8 s and then at  $10^\circ$  for 2.0 s, and analysed with FREALIGN<sup>42</sup>.

**Image processing and three-dimensional reconstruction.** An initial data set for complex I in DDM was obtained by manually picking particles with XIMDISP<sup>43</sup>. Seven-thousand six-hundred and thirty particles, from 366 micrographs, were used to generate initial maps in EMAN2 (ref. 44). Particles were boxed in  $280 \times 280$  pixels and contrast transfer function (CTF) parameters estimated internally. Reference-free classification was performed using the default EMAN2 parameters, and classes with distinct orientations selected (see Extended Data Fig. 1b for an example) to build initial models. The initial model that best matched the class averages was selected and two cycles of refinement performed in EMAN2. Subsequently, comparison of the model with the structure of complex I from *T. thermophilus* (Protein Data Bank (PDB) accession 4HEA<sup>4</sup>) suggested that it had the wrong hand; this observation was verified using tilt-pair analysis<sup>45</sup> and corrected (Extended Data Fig. 2). All further refinements were performed in RELION<sup>14</sup>, starting with maps that were low-pass filtered to 60 Å.

Typical micrographs prepared from complex I in Cymal 7 exhibited, on average, twice as many particles ( $\sim 40$  per micrograph) than those from complex I in DDM, so a larger data set was collected using Cymal 7. The reason for the difference in particle distribution is not clear—it may simply be a product of the grid preparation and freezing protocols. The class averages were used as a reference to pick particles automatically using RELION but many false positives were included, so all the images were inspected manually and particles too close to each other, aggregates and ice contaminants were deleted. The final data set contained 45,618 particles from 1,154 micrographs. The CTF was determined with CTFIND3 (ref. 46) using the images summed from all 72 frames. Subsequently, refinement was performed using frames 1–32 of each image (the last 40 frames were discarded), to produce a map with resolution of 5.86 Å and orientational accuracy of  $1.2^\circ$ . To check for overfitting, phases were randomized beyond 10 Å on individual images (frames 1–32), followed by refinement as for the normal images<sup>47</sup>. The results clearly show the presence of information beyond 10 Å (Extended Data Fig. 2). Note that RELION divides the data set into two halves at the initial step and calculates the resolution using a gold-standard Fourier shell correlation (FSC), so the phase randomization procedure serves here only as an additional control.

Modelling of the beam-induced movement of the complex I particles (using a running average of 11 movie frames in RELION) provided a modest improvement in resolution to 5.16 Å. The parameters from this analysis were then used to carry out a per-frame reconstruction in RELION (particle-polish), and a *B*-factor weighting was applied to each frame, resulting in a 4.8 Å map. The *B*-factor-weighted particles were subjected to 25 iterations of three-dimensional classification into 4 classes; this separated a major class containing 55% of the particles from smaller classes of 24, 14 and 6%. Difference maps revealed minor localized variations in some of the peripheral regions of the molecule, but no large-scale conformational variation was observed. The major class with 25,492 particles was refined and, after post-processing with RELION, a shape-mask, correction for the modulation transfer function (MTF) of the detector and a *B*-factor of  $-152$  (ref. 48) were applied, and filtered to 4.95 Å resolution (Extended Data Fig. 2C, note that the magnification and CTF values have not been refined). Despite containing a lower number of particles, the maps from this major class and the whole data set were comparable. Analysis of the local resolution by ResMap<sup>49</sup> showed that the core sections of the molecule, particularly the TMHs, have higher resolutions than the peripheral sections, and that (as expected) the detergent-phospholipid belt is at lower resolution (Extended Data Fig. 2e).

**Model building.** Model building was performed using Coot<sup>50</sup>. All the models described have been built as polyaniline chains, except for the residues that coordinate the FeS clusters. Note that the present model has not been refined, so it inevitably contains some errors and inaccuracies. Examples of the model fitted to the electron density are shown in Extended Data Fig. 3, and figures were created using the

PyMOL Molecular Graphics System or UCSF Chimera (<http://www.cgl.ucsf.edu/chimera/>).

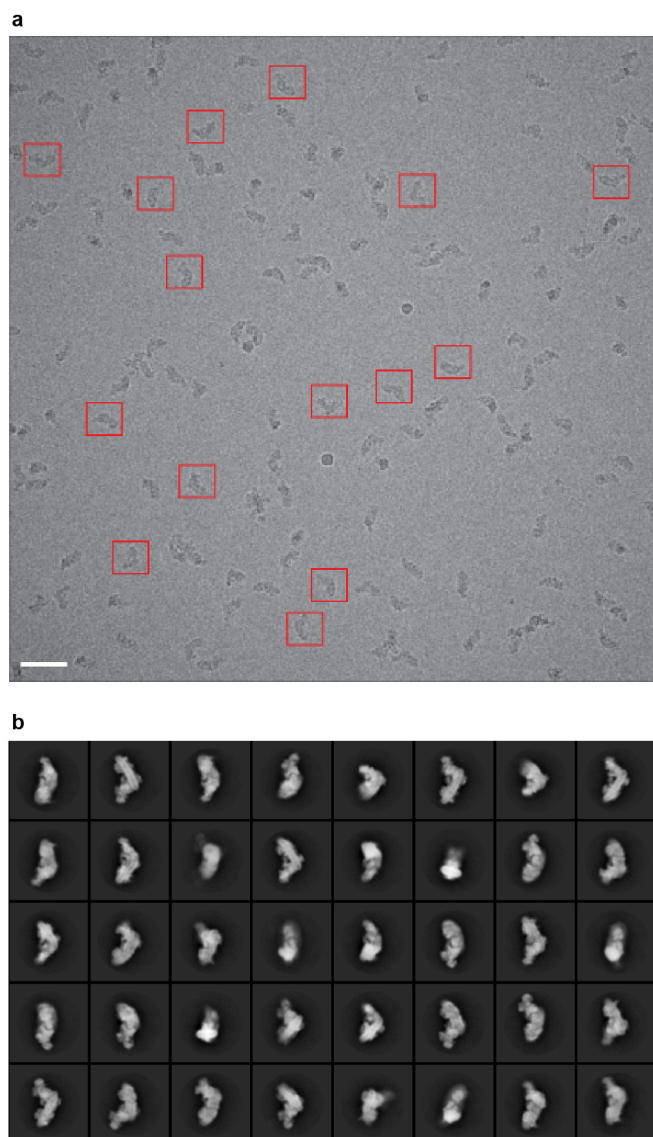
The seven core hydrophilic subunits of *B. taurus* complex I were modelled initially using the coordinates of *T. thermophilus* complex I (PDB accession 4HEA<sup>4</sup>) as a template, trimmed where the densities were ambiguous, and adjusted manually. FeS clusters were located using the highest peak densities in the unsharpened map, and the subunits were built around them. The 24 and 51 kDa subunits were easily built as they are well conserved and the connectivity in the densities is clearly resolved. The 49 kDa subunit has dominant secondary structures and, except for the N-terminal peptide, could be completely traced. Similarly, the PSST and TYKY subunits, with three FeS clusters, were readily built. The central part of the 30 kDa subunit, containing a mixture of  $\alpha$ -helices and  $\beta$ -strands, could be traced, but the path of the long unstructured C terminus is unclear. The 75 kDa subunit is the least conserved hydrophilic core subunit. The small domain containing the three FeS clusters is well resolved and could be traced easily using the *T. thermophilus* model, but considerable portions of the large, peripheral domain have poor density, low secondary structure content and low sequence similarity to *T. thermophilus*, and so could not be traced confidently. The seven core subunits in the membrane domain could all be readily traced, except for a few loop regions, and assigned using their similarity to the *T. thermophilus* subunits. The long transverse helix at the C terminus of subunit ND5 is well ordered and extends over ND4 and ND2. In better resolved regions of the map, protruding densities that are likely to be side chains of aromatic residues are observed (see Extended Data Fig. 3), and these features, along with secondary structure information and sequence alignments, were used to produce an optimized assignment for the *B. taurus* residue numbers in the modelled subunits (Extended Data Table 2), for use as a guide to the positions of individual residues.

Once the electron density for the core subunits had been assigned, models for the TMHs of the supernumerary subunits were built. A total of 18 TMHs were modelled, and when the density was clear they were extended. Connectivity was observed between four TMHs adjacent to subunit ND4 so they were combined into a single chain. To aid in supernumerary subunit assignments, the secondary structure of each subunit was predicted using PSIPRED<sup>51</sup> and TMHs were predicted using TMHMM2 (ref. 52), HMMTOP2 (ref. 53) and the TOPCONS suite<sup>54</sup> (seven methods in total) (Extended Data Table 4). Known structures of soluble proteins with high homology to the complex I supernumerary subunits were identified by HHpred<sup>55</sup> and used to build homology models in Modeller<sup>56</sup> and SwissModel<sup>57</sup> (Extended Data Table 5). Regions of the density map with features corresponding to the predicted structures were located manually. Long loop regions were trimmed, then the models were placed in the density, jiggle fit in Coot was used to find the best fit, and the models were adjusted manually. Finally, several additional tubular densities in the map were built as  $\alpha$ -helices. Most of them are located close to TMHs from the supernumerary subunits, but the connectivity to them is not clear; a higher resolution map will be necessary to assign these helices to their respective subunits.

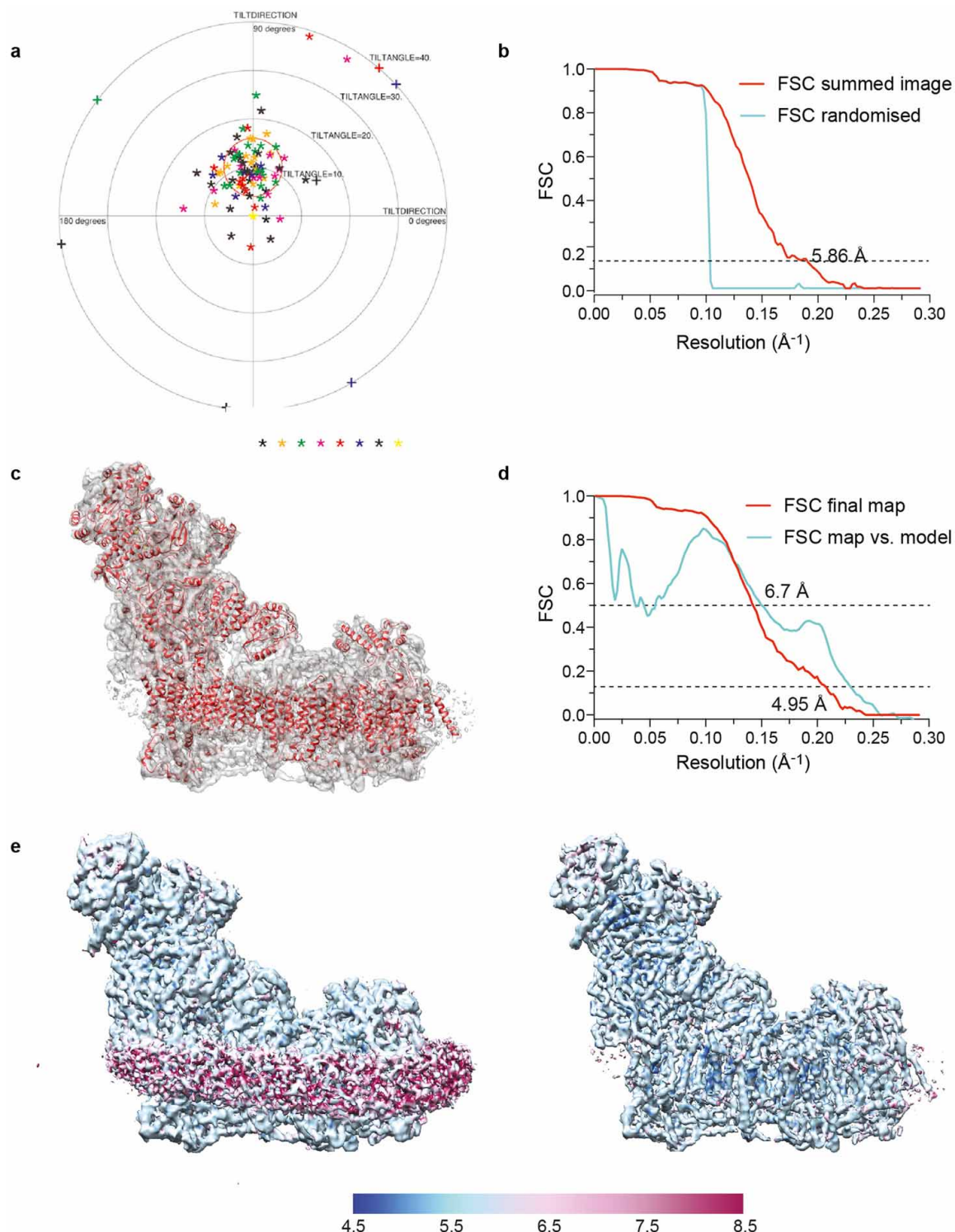
- Bellare, J. R., Davis, H. T., Scriven, L. E. & Talmon, Y. Controlled environment vitrification system: an improved sample preparation technique. *J. Electron Microsc. Tech.* **10**, 87–111 (1988).
- Grigorieff, N. FREALIGN: high-resolution refinement of single particle structures. *J. Struct. Biol.* **157**, 117–125 (2007).
- Smith, J. M. XIMDISP—a visualization tool to aid structure determination from electron microscope images. *J. Struct. Biol.* **125**, 223–228 (1999).
- Tang, G. et al. EMAN2: an extensible image processing suite for electron microscopy. *J. Struct. Biol.* **157**, 38–46 (2007).
- Henderson, R. et al. Tilt-pair analysis of images from a range of different specimens in single-particle electron cryomicroscopy. *J. Mol. Biol.* **413**, 1028–1046 (2011).
- Mindell, J. A. & Grigorieff, N. Accurate determination of local defocus and specimen tilt in electron microscopy. *J. Struct. Biol.* **142**, 334–347 (2003).
- Chen, S. et al. High-resolution noise substitution to measure overfitting and validate resolution in 3D structure determination by single particle electron cryomicroscopy. *Ultramicroscopy* **135**, 24–35 (2013).
- Rosenthal, P. B. & Henderson, R. Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. *J. Mol. Biol.* **333**, 721–745 (2003).
- Kucukelbir, A., Sigworth, F. J. & Tagare, H. D. Quantifying the local resolution of cryo-EM density maps. *Nature Methods* **11**, 63–65 (2014).
- Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).
- Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292**, 195–202 (1999).
- Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **305**, 567–580 (2001).
- Tusnády, G. E. & Simon, I. Principles governing amino acid composition of integral membrane proteins: applications to topology prediction. *J. Mol. Biol.* **283**, 489–506 (1998).



54. Bernsel, A., Viklund, H., Hennerdal, A. & Elofsson, A. TOPCONS: consensus prediction of membrane protein topology. *Nucleic Acids Res.* **37**, W465–W468 (2009).
55. Söding, J., Biegert, A. & Lupas, A. N. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.* **33**, W244–W248 (2005).
56. Eswar, N. *et al.* Comparative protein structure modeling using Modeller. *Curr. Protoc. Bioinform.* Chapter 5, Unit 5.6 (2006).
57. Arnold, K., Bordoli, L., Kopp, J. & Schwede, T. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics* **22**, 195–201 (2006).
58. Efremov, R. G. & Sazanov, L. A. Respiratory complex I: 'steam engine' of the cell? *Curr. Opin. Struct. Biol.* **21**, 532–540 (2011).
59. Krissinel, E. & Henrick, K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr. D* **60**, 2256–2268 (2004).
60. Balsa, E. *et al.* NDUF44 is a subunit of complex IV of the mammalian electron transport chain. *Cell Metab.* **16**, 378–386 (2012).
61. Johansson, K. *et al.* Structural basis for substrate specificities of cellular deoxyribonucleoside kinases. *Nature Struct. Biol.* **8**, 616–620 (2001).
62. King, J. D. *et al.* Predicting protein function from structure - the roles of short-chain dehydrogenase/reductase enzymes in *Bordetella* O-antigen biosynthesis. *J. Mol. Biol.* **374**, 749–763 (2007).
63. Parris, K. D. *et al.* Crystal structures of substrate binding to *Bacillus subtilis* holo-(acyl carrier protein) synthase reveal a novel trimeric arrangement of molecules resulting in three active sites. *Structure* **8**, 883–895 (2000).



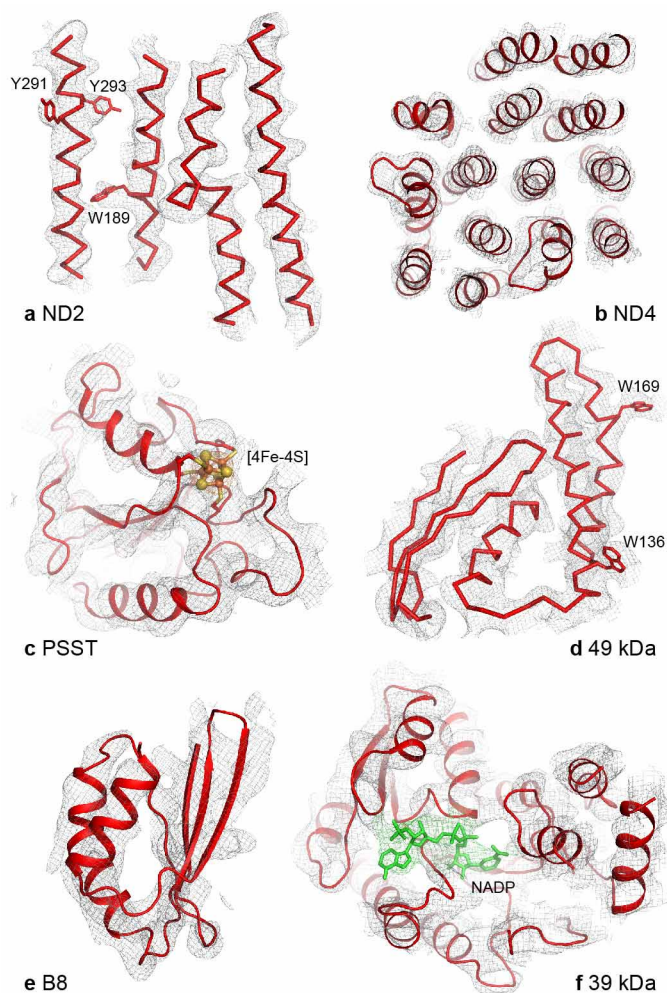
**Extended Data Figure 1 | Single-particle cryo-EM analysis of *B. taurus* complex I.** **a**, Typical micrograph of complex I particles imaged after freezing in vitreous ice on a holey-carbon grid. Some of the selected particles are marked with red boxes. Scale bar, 50 nm. **b**, Two-dimensional reference classification showing particles lying in different orientations in the ice. The size of each box is 280 pixels and the two-dimensional classification was made in RELION<sup>14</sup>.



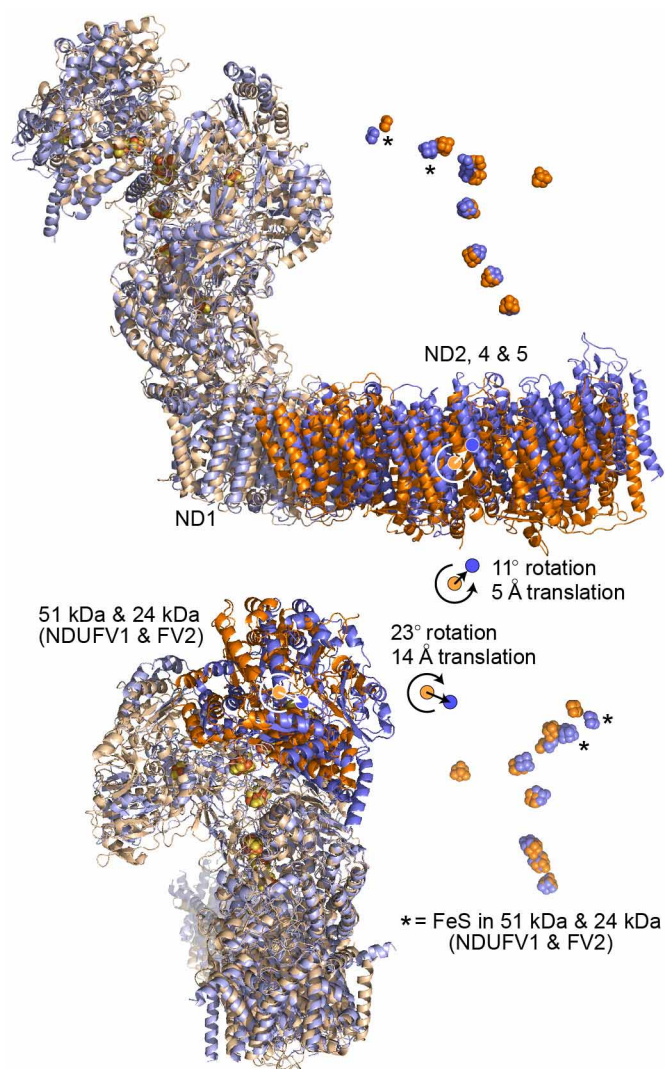
**Extended Data Figure 2 | Validation of the map and resolution.** **a**, Tilt-pair analysis<sup>45</sup> of complex I in Cymal-7. One-hundred complex I particles from eight image pairs, recorded with a relative tilt angle of 10°, were extracted and subjected to tilt-pair analysis with FREALIGN<sup>42</sup>. The outer radius of the plot is 40° and the orange circle centred at the expected tilt angle has a radius of 6°. **b**, Phase randomization to check for overfitting. Phases that are beyond 10 Å in each of the micrographs used in the final data set (frames 1–32) were randomized, and then refinement was performed as for a normal data set (FSC summed image corresponding to frames 1–32). As expected, the graph shows a drop in the Fourier shell correlation (FSC) curve at 10 Å, validating the presence of information beyond 10 Å in the images. Note that the use of gold-standard refinement procedures in RELION<sup>14</sup> prevents any overfitting, and this test was done only as an additional control. **c**, An overview of the final

map and the model built into it. **d**, FSC curves of the final map and of the model versus the map. The curve in red is the gold-standard FSC of the final map (after classification) and the resolution at FSC = 0.143 is ~4.95 Å. The curve in cyan is the FSC between the final map and the model, and at FSC = 0.5 the resolution is 6.7 Å. Note that the present model is not complete since it is only a polyaniline model without any side chains, and loop regions in a number of subunits have not been modelled. **e**, The final map of mammalian complex I was analysed with ResMap<sup>49</sup>. The left-hand panel (with lower density threshold) shows that the detergent-phospholipid belt is of lower resolution, and most of the protein regions of the map show resolution distributed from 5 to 6 Å. In the right-hand panel the map is shown at a higher density threshold, so the detergent-phospholipid belt is not visualized. Some of the interior parts of the map have resolution of 4.8–5 Å.



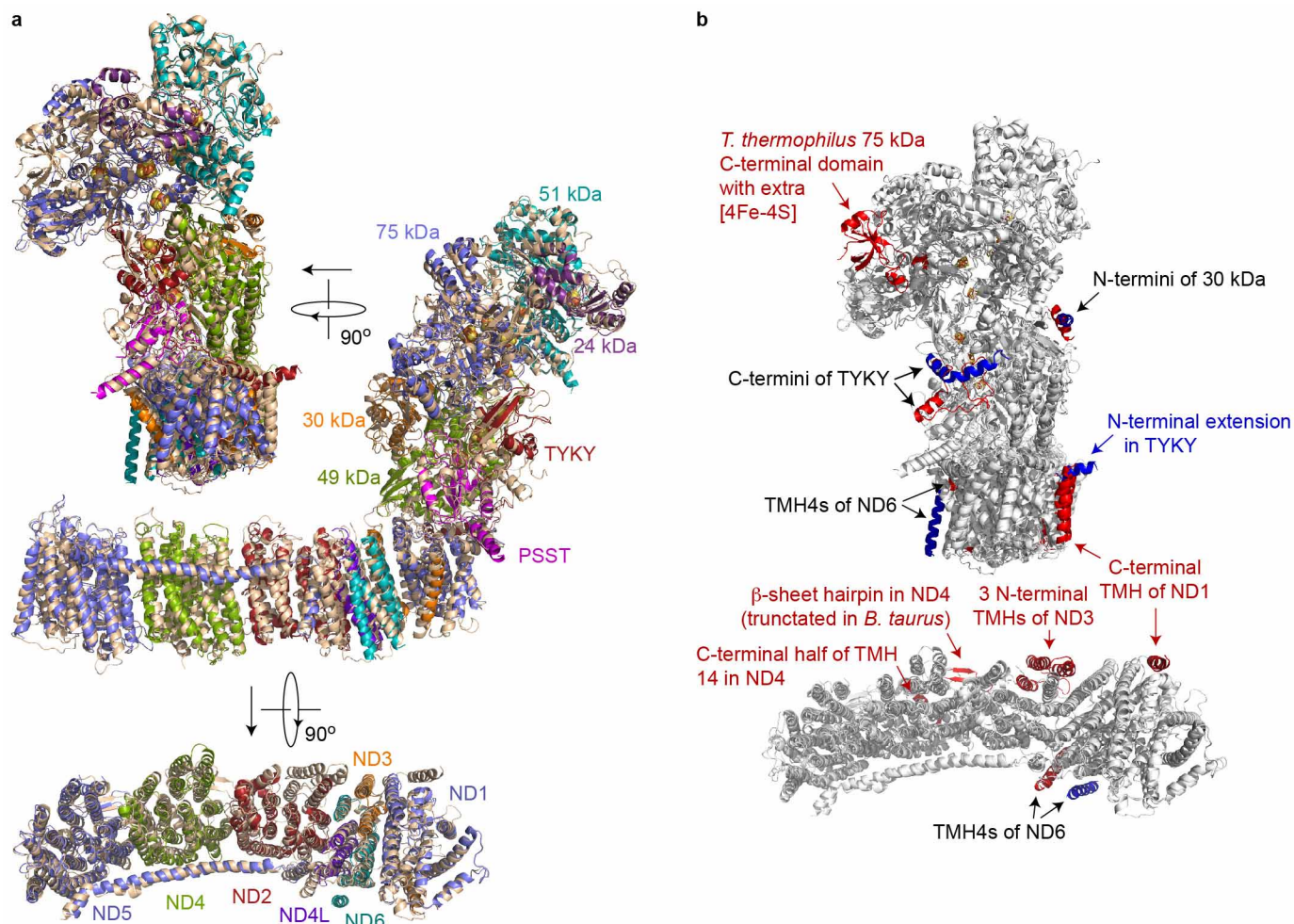


**Extended Data Figure 3 | Example regions of the density map with the model fitted to the map.** **a**, ND2 is shown from the membrane plane, highlighting the densities for three aromatic side chains and one of the helix-breaking loops. **b**, Subunit ND4 viewed from the matrix. **c**, The density for a [4Fe-4S] cluster and surrounding protein is shown in the PSST subunit. **d**, A region of the 49 kDa subunit shows a well resolved  $\alpha$ -helical stretch and aromatic side chains, and the  $\beta$ -strands are beginning to be resolved. **e**, Subunit B8 is an example of a supernumerary subunit in a peripheral region of the molecule. **f**, Density consistent with a bound nucleotide is observed in the 39 kDa subunit, in a similar position to in homologous structures and as expected from analysis of *Y. lipolytica* complex I (ref. 27). However, the present resolution of the map precludes the inclusion of this nucleotide in the final model.



#### Extended Data Figure 4 | Global comparison of the core subunit structures of bacterial and mammalian complex I.

The core subunits from *B. taurus* are in blue, and from *T. thermophilus* (PDB accession 4HEA<sup>4</sup>) in orange. The structures have been superimposed using ND1 (the heel subunit). Top: the ND2, ND4 and ND5 domain is rotated in *B. taurus* relative to in *T. thermophilus*, increasing the curvature in the *B. taurus* membrane domain. The complex is viewed along the 11° rotation vector (orange) that maps the *T. thermophilus* ND2, ND4 and ND5 domain to the *B. taurus* domain, along with a small 5 Å translation to superimpose the domain centres. Correspondingly, the ND3, ND4L and ND6 domains are superimposed by a 4° rotation and a 1 Å translation. Rotation of ND2, 4 and 5 about the long axis of the domain, as noted for *Y. lipolytica*<sup>58</sup>, is not observed. Bottom: the NADH dehydrogenase domain containing the 51 and 24 kDa subunits is rotated by 23° and translated by 14 Å in *B. taurus*, relative to in *T. thermophilus*, causing the FeS chains to diverge as the distance from ND1 increases. A similar rotation was observed in *Y. lipolytica*<sup>58</sup>. The complex is viewed from behind ND1. Correspondingly, the 49 kDa, PSST and TYKY subunits are superimposed by a 6° rotation and a 2 Å translation. The structures were analysed using Superpose from the CCP4 suite<sup>59</sup> and the 75 kDa and 30 kDa subunits were not included due to their lower structural conservation.



**Extended Data Figure 5 | Comparison of the individual structures of the core subunits of bacterial and mammalian complex I.** **a**, The structure of each subunit from *T. thermophilus* (wheat) (PDB accession 4HEA<sup>4</sup>) has been superimposed separately on its corresponding subunit from *B. taurus* (coloured as labelled) with the transverse helix plus TMH16 of ND5 also aligned separately. The complexes are viewed from behind ND1 (top), from the side (middle) and from the matrix (bottom, ND subunits only). **b**, Observed

differences in the structures of the core subunits of *B. taurus* and *T. thermophilus* complexes I. Grey, conserved structure from *B. taurus* and *T. thermophilus* (PDB accession 4HEA<sup>4</sup>); red, structural elements present only in *T. thermophilus*; blue, structural elements present only in *B. taurus*. The C-terminal domain of the 75 kDa subunit is not resolved in *B. taurus*, but its structure is clearly different to in *T. thermophilus*.



Extended Data Table 1 | Reference table for the nomenclature of the core subunits of complex I

Domain	Chain identifier	<i>Bos taurus</i>	<i>Homo sapiens</i>	<i>Yarrowia lipolytica</i>	<i>Thermus thermophilus</i>	<i>Escherichia coli</i>
Hydrophilic domain	G	75 kDa	NDUFS1	NUAM	Nqo3	NuoG
	F	51 kDa	NDUFV1	NUBM	Nqo1	NuoF
	D	49 kDa	NDUFS2	NUCM	Nqo4	NuoCD
	C	30 kDa	NDUFS3	NUGM	Nqo5	
	E	24 kDa	NDUFV2	NUHM	Nqo2	NuoE
	B	PSST	NDUFS7	NUKM	Nqo6	NuoB
	I	TYKY	NDUFS8	NUIM	Nqo9	NuoI
Membrane domain	H	ND1	ND1	NU1M	Nqo8	NuoH
	N	ND2	ND2	NU2M	Nqo14	NuoN
	A	ND3	ND3	NU3M	Nqo7	NuoA
	M	ND4	ND4	NU4M	Nqo13	NuoM
	L	ND5	ND5	NU5M	Nqo12	NuoL
	J	ND6	ND6	NU6M	Nqo10	NuoJ
	K	ND4L	ND4L	NULM	Nqo11	NuoK

In the text the names of the subunits from *B. taurus* are used, with the names from the human enzyme presented alongside as appropriate.

Extended Data Table 2 | Summary of the models of the core subunits of *B. taurus* complex I

Subunit	Total residues *	Modelled residues	Poorly resolved / uncertain residue numbering	Unresolved residues	Unresolved elements (>10 residues)	%Modelled	%Identity†	r.m.s.d. †
ND1	318	3 - 200 219 - 242 253 - 315		1 - 2 201 - 218 243 - 252 316 - 318	Matrix loop (TMH 5 - 6) IMS loop (TMH 6 - 7)	90% (285/318)	42% (132/318)	1.60 Å
ND2	347	2 - 300 320 - 346	TMH11	1 301 - 319 347	Matrix loop (TMH 10 - 11)	94% (326/347)	25% (86/347)	2.08 Å
ND3	115	2 - 23 52 - 112		1 24 - 51 113 - 115	Matrix loop (TMH 1 - 2)	72% (83/115)	27% (31/115)	2.05 Å
ND4	459	3 - 415 430 - 455	TMH14	1 - 2 416 - 429 456 - 459	Matrix loop (TMH 13 - 14)	96% (439/459)	24% (111/459)	2.20 Å
ND4L	98	1 - 84		85 - 98	Matrix loop (C-terminus)	86% (84/98)	21% (21/98)	2.66 Å
ND5	606	4 - 22 28 - 358 363 - 400 408 - 466 487 - 513 520 - 604	TMH1  TMH13 & TMH14 TMH15 Transverse helix & TMH16	1-3 23-27 359-362 401-407 467-486 514-519 605-606	Matrix loop (TMH 1-2) Matrix loop (TMH 11-12) IMS loop (TMH 12-13) IMS loop (TMH 14-15) TMH15 to transverse helix	92% (558/606)	31% (187/606)	2.53 Å
ND6	175	2 - 76 85 - 107 140 - 172	TMH5	1 77 - 84 108 - 139 173 - 175	Matrix loop (TMH 3 - 4) IMS loop (TMH 4 - 5)	75% (131/175)	16% (28/175)	1.83 Å
75 kDa NDUFS1	704	8 - 125 136 - 318 326 - 347 367 - 400 404 - 410 425 - 495 525 - 530 542 - 627	The large domain (222 - 704) is generally poorly resolved. The sequence alignment is weak and the secondary structure content low. Residues 404 - 629 are particularly poorly resolved.	1 - 7 126 - 135 319 - 325 348 - 366 400 - 403 411 - 424 496 - 524 531 - 541 628 - 704	Probable loop region  Probable loop region Probable loop region Probable loop region Probable subdomain	75% (527/704)	27% (189/704)	1.96 Å  1 - 221: 1.57 Å  222 - 704: 2.11 Å
51 kDa NDUFV1	444	31 - 441	Flavin and NADH binding site (63 - 72, 99 - 104, 181 - 189, 300 - 304, 327 - 333)	1 - 30 442 - 444	N-terminal peptide	93% (411/444)	43% (191/444)	1.61 Å
49 kDa NDUFS2	430	47 - 430	3-strand $\beta$ -sheet (47 - 79)	1 - 46	N-terminal region	89% (384/430)	42% (179/430)	1.41 Å
30 kDa NDUFS3	228	15 - 168	Numbering uncertain to 72 Loop / $\beta$ -strand (73 - 83)	1 - 14 169 - 228	N-terminal peptide C-terminal region	68% (154/228)	24% (54/228)	1.66 Å
24 kDa NDUFV2	217	20 - 178	Loop 126 - 132	1 - 19 179 - 217	N-terminal peptide C-terminal region	73% (159/217)	27% (59/217)	1.57 Å
PSST NDUFS7	179	27 - 169	Loop 68 - 79	1 - 26 170 - 179	N-terminal peptide	80% (143/179)	49% (88/179)	1.44 Å
TYKY NDUFS8	176	15 - 176		1 - 14	N-terminal peptide	92% (162/176)	36% (63/176)	1.89 Å

\* For proteins with a mitochondrial-targeting pre-sequence, residue 1 is the first residue of the mature protein<sup>2,3</sup>.† The percentage identity and the root mean squared deviation (r.m.s.d., calculated using PDBeFOLD<sup>59</sup>) are between the sequences and structures of the subunits of *B. taurus* and *T. thermophilus* (PDB accession 4HEA) complex I.

Extended Data Table 3 | Distances between the redox cofactors in structural models of complex I

Cofactors*	<i>T. thermophilus</i>				<i>B. taurus</i>	
	hydrophilic domain (2FUG.pdb <sup>7</sup> )		complex I (4HEA.pdb <sup>4</sup> )		complex I (this work)	
	centre <sup>†</sup>	edge <sup>†</sup>	centre <sup>†</sup>	edge <sup>†</sup>	centre <sup>†</sup>	edge <sup>†</sup>
N1a - Flavin	15.4	12.3	15.9	13.1	15.9 <sup>‡</sup>	13.1 <sup>‡</sup>
Flavin - cluster 1 (N3)	12.5	7.6	12.2	7.3	12.2 <sup>‡</sup>	7.2 <sup>‡</sup>
N1a - cluster 1 (N3)	22.1	19.4	22.3	19.7	21.1	18.0
Cluster 1 (N3) - cluster 2	14.0	11.0	13.7	10.7	14.0	11.0
Cluster 1 (N3) - cluster 3	17.4	13.8	17.1	13.4	18.4	14.5
Cluster 2 - cluster 3	13.5	10.7	13.0	9.9	12.7	9.7
Cluster 3 - cluster 4	12.2	8.5	12.4	8.6	12.8	8.7
Cluster 4 - cluster 5	16.8	14.0	16.5	13.6	16.8	14.0
Cluster 5 - cluster 6	12.1	9.4	12.1	9.3	12.1	9.3
Cluster 6 - cluster 7 (N2)	13.7	10.5	13.5	10.2	13.6	10.5
Cluster 1 (N3) - cluster 7 (N2)	61.1	57.6	60.5	57.0	61.5	58.1

\*The [2Fe–2S] cluster in the 24 kDa subunit (known as N1a) is on the other side of the flavin from the main cofactor chain. The [4Fe–4S] cluster in the 51 kDa subunit (known as N3) is the first cluster in the chain and the [4Fe–4S] cluster in subunit PSST (known as N2) is the last (seventh) cluster in the chain.

<sup>†</sup>The distances are in Å, between the geometric centres of the Fe and S cluster cores or the flavin isoalloxazine ring system (centre), or between the centres of the two closest atoms (edge) as commonly used in calculations of electron transfer rates. Distances are estimated to be accurate to within 1 Å.

<sup>‡</sup>The position of the flavin in *B. taurus* is poorly resolved and has been approximated using its position in PDB accession 4HEA.



Extended Data Table 4 | Knowledge about the supernumerary subunits of *B. taurus* complex I

<i>B. taurus</i> subunit*	<i>H. sapiens</i> subunit*	Subcomplex <sup>†</sup>	Sequence information	Predicted TMHs <sup>‡</sup>
10 kDa	NDUFV3	I $\alpha$ and I $\lambda$ .		0
18 kDa	NDUFS4	I $\alpha$ and I $\lambda$ .		0
15 kDa	NDUFS5	I $\alpha$ only	CX <sub>9</sub> C motif, intermembrane space <sup>37</sup>	0
13 kDa	NDUFS6	I $\alpha$ and I $\lambda$ .	PFAM zinc-finger motif CX <sub>8</sub> HX <sub>15</sub> CX <sub>2</sub> C	0
MWFE	NDUFA1	I $\alpha$ only		1
B8	NDUFA2	I $\alpha$ and I $\lambda$ .		0
B9	NDUFA3	I $\alpha$ only		1
B13	NDUFA5	I $\alpha$ and I $\lambda$ .		0
B14	NDUFA6	I $\alpha$ only	LYR motif <sup>32</sup>	0
B14.5a	NDUFA7	I $\alpha$ and I $\lambda$ .		0
PGIV	NDUFA8	I $\alpha$ only	Two CX <sub>9</sub> C motifs, PFAM CHCH domain intermembrane space <sup>37</sup>	0
39 kDa	NDUFA9	I $\alpha$ only	Short-chain dehydrogenase reductase family, NADP binding <sup>26,27</sup>	0
42 kDa	NDUFA10	I $\alpha$ only (low level)	Similarity to deoxynucleoside kinases <sup>24</sup>	0
B14.7	NDUFA11	I $\alpha$ (I $\lambda$ at low level)		3 or 4
B17.2	NDUFA12	I $\alpha$ and I $\lambda$ .		0
B16.6	NDUFA13	I $\alpha$ and I $\lambda$ .		1
SDAP	NDUFAB1	both I $\alpha$ and I $\beta$	Acyl-carrier protein <sup>29,30</sup>	0
MNLL	NDUFB1	I $\beta$		0 (or 1)
AGGG	NDUFB2	I $\beta$		1 (or 0)
B12	NDUFB3	I $\beta$		1
B15	NDUFB4	both I $\alpha$ and I $\beta$		1
SGDH	NDUFB5	I $\beta$		1
B17	NDUFB6	I $\beta$		1
B18	NDUFB7	I $\beta$	CX <sub>9</sub> C motif, intermembrane space <sup>37</sup>	0
ASHI	NDUFB8	I $\beta$		1
B22	NDUFB9	I $\beta$	LYR motif <sup>32</sup>	0
PDSW	NDUFB10	I $\beta$		0
ESSS	NDUFB11	I $\beta$		1
KFYI	NDUFC1	none		1
B14.5b	NDUFC2	I $\beta$ (low level)		1 or 2

\* The former subunit MLRQ (NDUFA4) is no longer considered a subunit of complex I (ref. 60).

<sup>†</sup> Subcomplex I $\lambda$ , which contains the 7 hydrophilic core subunits and 8–9 supernumerary subunits, represents a considerable portion of the hydrophilic domain of complex I. Subcomplex I $\alpha$ , which contains all the subunits of subcomplex I $\lambda$ , plus core subunit ND6 and 9–10 additional supernumerary subunits, represents the hydrophilic domain of complex I plus associated membrane subunits. Subcomplex I $\beta$ , which contains ND4 and ND5 and 12–13 supernumerary subunits, represents part of the membrane domain<sup>3</sup>.

<sup>‡</sup> TMHs were predicted using TMHMM2 (ref. 52), HMMTOP2 (ref. 53) and the TOPCONS suite<sup>54</sup> (seven methods in total) and are presented as consensus values with less represented values in brackets and single outliers discarded.

Extended Data Table 5 | Summary of the models of the supernumerary subunits of *B. taurus* complex I

Subunit	Chain identifier	Total residues*	PDB model†	Aligned residues	%Identity	Modelled residues	%Modelled	r.m.s.d.‡
42 kDa§ NDUFA10	O	320	2OCP <sup>61</sup>	21 - 252	21% (49/232)	22 - 54 79 - 167 172 - 210 222 - 241	57% (181/320)	1.91 Å
39 kDa NDUFA9	P	345	2Q1W <sup>62</sup>	19 - 325	13% (41/307)	19 - 185 203 - 250 285 - 321	73% (252/345)	2.52 Å
18 kDa§ NDUFS4	Q	133	2JYA	33 - 133	37% (37/101)	33 - 59 76 - 116	52% (69/133)	2.42 Å
13 kDa§ NDUFS6	R	96	2JRR	44 - 96	34% (18/53)	47 - 93	49% (47/96)	1.97 Å
B8 NDUFA2	S	99	1S3A <sup>20</sup>	1 - 99	94% (93/99)	17 - 96	81% (80/99)	2.18 Å
SDAP- $\alpha$ NDUFAB1	T	88	1F80 <sup>63</sup>	8 - 84	36% (28/77)	9 - 23 28 - 82	81% (71/88)	1.18 Å
SDAP- $\beta$ NDUFAB1	U	88	1F80 <sup>63</sup>	8 - 84	36% (28/77)	8 - 82	85% (75/88)	1.36 Å
B13§ NDUFA5	V	116				1 - 71¶	61% (71/116)	
B14 NDUFA6	W	128				1 - 72¶	56% (72/128)	
PGIV NDUFA8	X	172	2LQL <sup>36</sup>	35 - 114	23% (18/80)	1 - 80¶	46% (79/172)	2.40 Å
B14.7 NDUFA11	Y	141				1 - 106¶	75% (106/141)	
B16.6 NDUFA13	Z	144				33 - 97	45% (65/144)	
B9§ NDUFA3 or MWFE§ NDUFA1	a	154				1 - 29¶	46% (71/154)	
	b					1 - 42¶		

\* For proteins with a mitochondrial-targeting pre-sequence, residue 1 is the first residue of the mature protein<sup>2,3</sup>.

† Known structures with high homology to the complex I subunits<sup>20, 36, 61-63</sup> were identified by HHpred<sup>55</sup>.

‡ The percentage identity and the r.m.s.d. (calculated using PDBFOLD<sup>59</sup>) are between the sequences and structures of the subunits of *B. taurus* complex I and the PDB models.

§ Subunit with less certain assignment.

¶ Residue numbers are arbitrary and not assigned to the sequence.

# Turbulent heating in galaxy clusters brightest in X-rays

I. Zhuravleva<sup>1,2</sup>, E. Churazov<sup>3,4</sup>, A. A. Schekochihin<sup>5,6</sup>, S. W. Allen<sup>1,2,7</sup>, P. Arévalo<sup>8,9</sup>, A. C. Fabian<sup>10</sup>, W. R. Forman<sup>11</sup>, J. S. Sanders<sup>12</sup>, A. Simionescu<sup>13</sup>, R. Sunyaev<sup>3,4</sup>, A. Vikhlinin<sup>11</sup> & N. Werner<sup>1,2</sup>

**The hot ( $10^7$  to  $10^8$  kelvin), X-ray-emitting intracluster medium (ICM) is the dominant baryonic constituent of clusters of galaxies. In the cores of many clusters, radiative energy losses from the ICM occur on timescales much shorter than the age of the system<sup>1–3</sup>. Unchecked, this cooling would lead to massive accumulations of cold gas and vigorous star formation<sup>4</sup>, in contradiction to observations<sup>5</sup>. Various sources of energy capable of compensating for these cooling losses have been proposed, the most promising being heating by the supermassive black holes in the central galaxies, through inflation of bubbles of relativistic plasma<sup>6–9</sup>. Regardless of the original source of energy, the question of how this energy is transferred to the ICM remains open. Here we present a plausible solution to this question based on deep X-ray data and a new data analysis method that enable us to evaluate directly the ICM heating rate from the dissipation of turbulence. We find that turbulent heating is sufficient to offset radiative cooling and indeed appears to balance it locally at each radius—it may therefore be the key element in resolving the gas cooling problem in cluster cores and, more universally, in the atmospheres of X-ray-emitting, gas-rich systems on scales from galaxy clusters to groups and elliptical galaxies.**

Perseus and Virgo (also known as M87) are well-studied, nearby, cool-core clusters of galaxies in which the central cooling times, owing to the emission of X-rays, are an order of magnitude shorter than the Hubble time (Methods and Extended Data Fig. 1). X-ray observations show that the ICM in central regions of these clusters is disturbed, suggesting that it might be turbulent. The most likely drivers of this turbulence are mechanically powerful active galactic nuclei (AGNs) in the central galaxies of both clusters, which inflate bubbles of relativistic plasma in the ICM. During the inflation and subsequent buoyant rise of these bubbles, internal waves and turbulent motion in the gas can be excited<sup>10–12</sup> and must eventually dissipate into heat. To determine whether this heating is sufficient to balance radiative losses and prevent net cooling, one must estimate the turbulent heating rate—and for that, a measurement is needed of the root mean squared turbulent velocity amplitude  $V$  as a function of length scale  $l$ . Then the turbulent heating rate in the gas with mass density  $\rho$  is (dimensionally)  $Q_{\text{turb}} \sim \rho V^3/l$  to within some constant of order unity that depends on the exact properties of the turbulent cascade. (We use a tilde between two quantities to indicate order-of-magnitude equivalence.)  $Q_{\text{turb}}$  has never previously been probed directly, mainly because of two difficulties. In this Letter, we propose ways of overcoming both, leading to an observational estimate of  $Q_{\text{turb}}$  and the tentative conclusion that it is sufficient to reheat the ICM.

The energy resolution of current X-ray observatories is insufficient to measure gas velocities in the ICM, or their dependence on scale. Here we circumvent this problem by instead measuring gas density fluctuations and inferring from their power spectrum the power spectrum of

the velocities. A simple theoretical argument, supported by numerical simulations, shows that in relaxed galaxy clusters, where the gas motions are subsonic, the root mean squared amplitudes of the density and one-component velocity fluctuations are proportional to each other on each scale  $l = k^{-1}$  within the inertial range<sup>13,14</sup>:  $\delta\rho_k/\rho_0 \approx \eta_1 V_{1,k}/c_s$ , where  $\rho_0$  is the mean gas density,  $c_s$  is the sound speed and  $\eta_1$  is the proportionality coefficient  $\sim 1$  set by gravity wave physics on large, buoyancy-dominated scales<sup>13</sup>. Here we define  $V_{1,k}$  by  $3V_{1,k}^2/2 = k_1 E(k_1)$ , where  $k_1 = 2\pi k$  is the Fourier wave number and  $E(k_1)$  is the energy spectrum of the three-dimensional velocity field;  $\delta\rho_k/\rho_0$  is defined analogously in terms of the density fluctuation spectrum, but without the factor of  $3/2$ . Unsharp-masked images of the Perseus cluster show ripple-like structures in the core, reminiscent of either sound waves<sup>15,16</sup> or stratified turbulence<sup>13,17</sup> (Methods). Here we investigate the consequences of the second of these possibilities (which may be more likely if the stirring of the ICM by the AGN ejecta is of sufficiently low frequency).

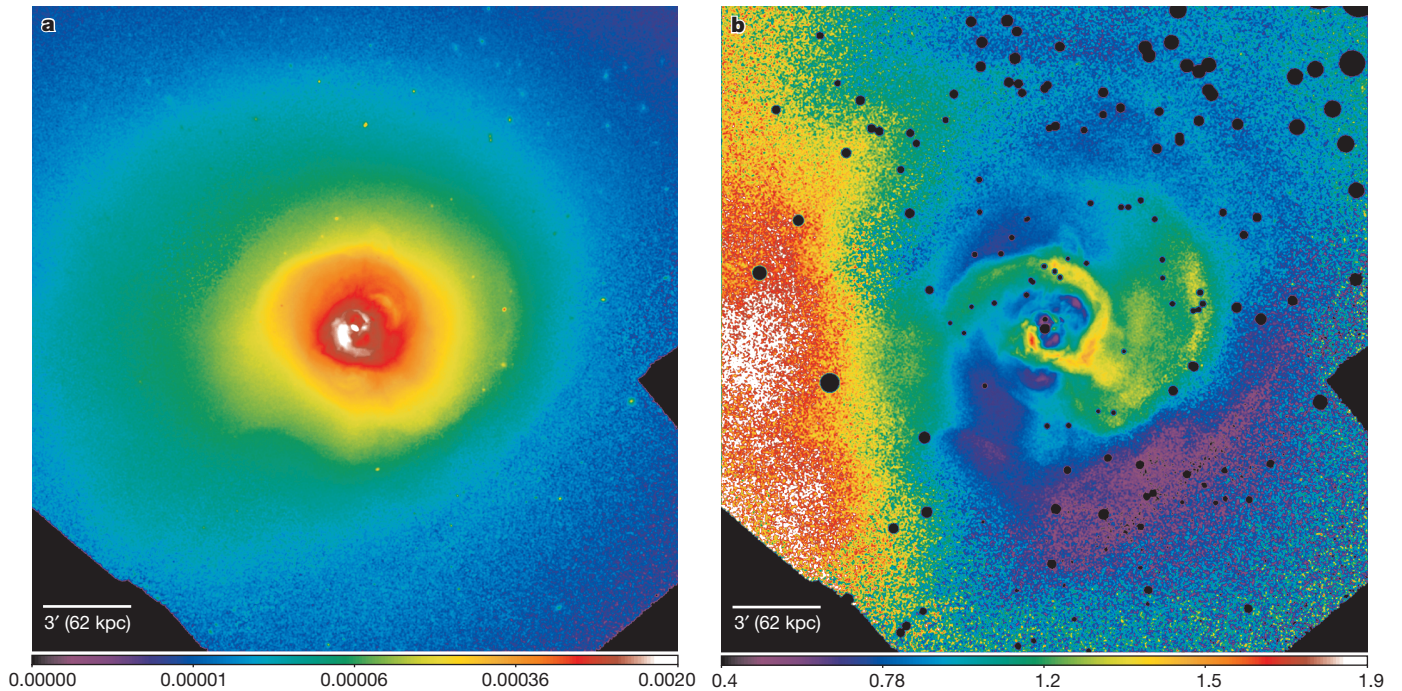
The high statistical precision obtained by Chandra in a 1.4 Ms observation of the Perseus cluster core makes this data set ideal for probing density structures over a range of spatial scales. Figure 1 shows the mosaic image and a residual image, made by dividing the mosaic image by a spherically symmetric  $\beta$ -model of the mean intensity profile with core radius  $1.26'$  (equivalent to  $\sim 26$  kpc at the distance of Perseus) and slope  $\beta = 0.53$  (Methods and Extended Data Fig. 2). Using the modified  $\Delta$ -variance method<sup>18</sup>, we calculate the power spectra of surface brightness fluctuations in a set of concentric annuli (Extended Data Fig. 3), each with width  $1.5'$  (31 kpc), and deduce from them the amplitudes of density fluctuations across a range of spatial scales. The typical  $\delta\rho_k/\rho_0$  at  $k^{-1} \sim 20$  kpc varies from  $\sim 20\%$  inside the central  $1.5'$  (31 kpc) to  $\sim 7\%$  at an angular distance of  $10.5'$  ( $\sim 218$  kpc) from the cluster centre (I.Z. *et al.*, manuscript in preparation). We have also performed a similar analysis for a  $\sim 600$  ks Chandra observation of the Virgo cluster.

Figure 2 shows examples of the velocity amplitudes  $V_{1,k}$  inferred from the density amplitudes  $\delta\rho_k/\rho_0$  using the relation  $\eta_1 V_{1,k}/c_s \approx \delta\rho_k/\rho_0$ , in two different annuli for each of the two clusters. In these examples, over the range of spatial scales where the measurements are least affected by systematic and statistical uncertainties,  $V_{1,k}$  varies from  $\sim 70$  km s<sup>−1</sup> to  $\sim 145$  km s<sup>−1</sup> in Perseus. In the full set of seven annuli from the centre to  $10.5'$  ( $\sim 218$  kpc), the range of velocities is larger, up to  $210$  km s<sup>−1</sup>. In Virgo, the typical velocity amplitudes in all annuli are smaller, between  $43$  and  $140$  km s<sup>−1</sup>, but the corresponding spatial scales are smaller too.

These (inferred) velocity spectra can be used to estimate the heating rate  $Q_{\text{turb}} \sim \rho V^3/l$ . The second difficulty mentioned earlier is that normally  $l$  here is taken to be the energy-containing scale of the turbulence, which is difficult to determine or even define unambiguously: in theory, several characteristic scales (for example the distance from the centre, various scale heights and the like) are present in the problem<sup>19</sup>. The

<sup>1</sup>Kavli Institute for Particle Astrophysics and Cosmology, Stanford University, 452 Lomita Mall, Stanford, California 94305-4085, USA. <sup>2</sup>Department of Physics, Stanford University, 382 Via Pueblo Mall, Stanford, California 94305-4060, USA. <sup>3</sup>Max Planck Institute for Astrophysics, Karl-Schwarzschild-Strasse 1, D-85741 Garching, Germany. <sup>4</sup>Space Research Institute (IKI), Profsoyuznaya 84/32, Moscow 117997, Russia. <sup>5</sup>Rudolf Peierls Centre for Theoretical Physics, University of Oxford, 1 Keble Rd, Oxford OX1 3NP, UK. <sup>6</sup>Merton College, University of Oxford, Merton St, Oxford OX1 4JD, UK. <sup>7</sup>SLAC National Accelerator Laboratory, 2575 Sand Hill Road, Menlo Park, California 94025, USA. <sup>8</sup>Instituto de Física y Astronomía, Facultad de Ciencias, Universidad de Valparaíso, Gran Bretaña N 1111, Playa Ancha, Valparaíso, Chile. <sup>9</sup>Instituto de Astrofísica, Facultad de Física, Pontificia Universidad Católica de Chile, 306, Santiago 22, Chile. <sup>10</sup>Institute of Astronomy, University of Cambridge, Madingley Road, Cambridge CB3 0HA, UK. <sup>11</sup>Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, Massachusetts 02138, USA. <sup>12</sup>Max-Planck-Institut für extraterrestrische Physik, Giessenbachstrasse 1, D-85748 Garching, Germany. <sup>13</sup>Japan Aerospace Exploration Agency, 3-1-1 Yoshinodai, Sagamihara, Kanagawa 252-5210, Japan.





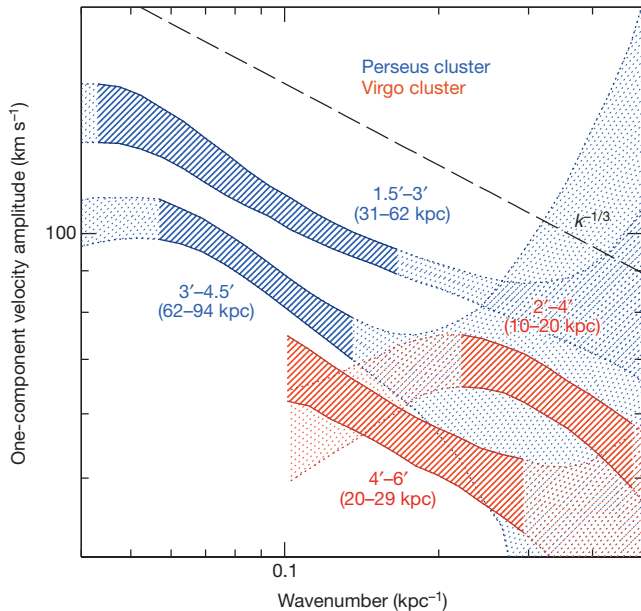
**Figure 1 | X-ray image of the core of the Perseus cluster.** **a**, X-ray surface brightness in units of counts per second per pixel (colour scale), obtained in the 0.5–3.5 keV energy band from Chandra observations. **b**, The same divided by the mean surface brightness profile, highlighting the relative density fluctuations. The images are smoothed with a 3'' Gaussian. Black circles:

excised point sources (Methods). The redshift is taken to be  $z = 0.01756$  (redshift of the central galaxy); hence, the angular diameter distance is 72 Mpc (for  $h = 0.72$ ,  $\Omega_m = 0.3$ ,  $\Omega_\Lambda = 0.7$ ) and 1'' corresponds to a length scale of 20.82 kpc.

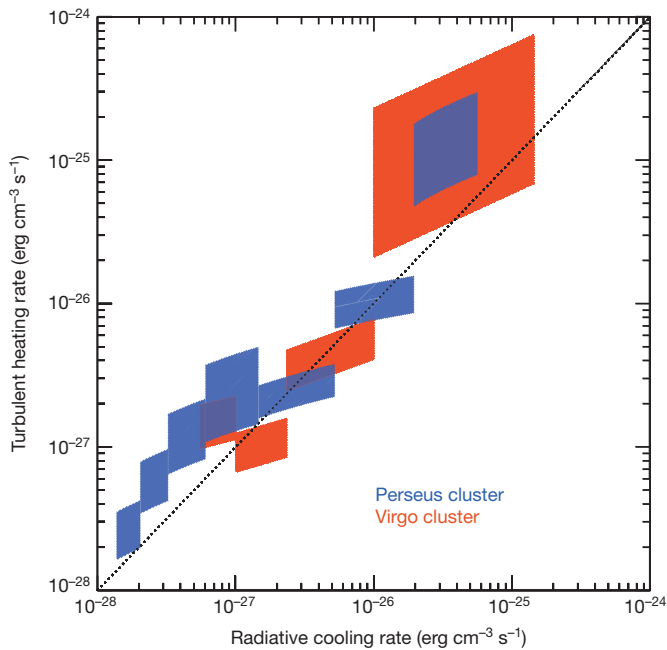
measured spectra (Fig. 2) do not necessarily offer clarity about the injection scale, because at low  $k$  they are dominated by large-scale inhomogeneities and the radial width of the chosen annuli. However, in a turbulent cascade, the energy spectrum in the inertial range should have a universal

form depending only on  $k$  and the mean, density-normalized dissipation rate  $\varepsilon = Q_{\text{turb}}/\rho_0$ . Assuming purely hydrodynamic<sup>20</sup> turbulence, the energy spectrum should be  $E(k_1) = C_K \varepsilon^{2/3} k_1^{-5/3}$ , where  $C_K \approx 1.65$  is the Kolmogorov constant<sup>21,22</sup>. The turbulent energy flux on any scale in the inertial range will be the same and equal to the mean dissipation rate: accounting for our conventions  $k = 1/l = k_1/2\pi$  and  $V_{1,k} = [2k_1 E(k_1)/3]^{1/2}$ , we obtain  $Q_{\text{turb}} = \rho_0 \varepsilon = C_Q \rho_0 V_{1,k}^3 k$ , where  $C_Q = 3^{3/2} 2\pi / (2C_K)^{3/2} \approx 5$  is a dimensionless constant whose value should be treated as a fiducial number. Indeed, although the constant-flux, Kolmogorov-like nature of the turbulence is probably a good assumption, the specific constant  $C_Q$  will depend on more detailed properties of the turbulent cascade (for example magnetohydrodynamic rather than hydrodynamic<sup>23</sup>) and, in particular, on the types of fluctuation (velocity, magnetic, density<sup>24</sup>) that carry the total injected energy flux to small scales. We will not be concerned here with a precise determination of  $C_Q$ . It is clearly an order-unity number and it is also clear that our estimate for the turbulent heating rate can be used only if we identify, for each of the annuli where we calculated  $V_{1,k}$ , a  $k$  interval in which  $V_{1,k}^3 k$  stays approximately constant in  $k$ . Remarkably, our measured velocities are indeed consistent with  $V_{1,k} \sim k^{-1/3}$ , accounting for the errors and uncertainties associated with finite resolution and with our method of extracting power spectra<sup>25</sup>.

Because of order-unity uncertainties in the determination of  $Q_{\text{turb}}$ , the question of the heating–cooling balance reduces to whether the local  $Q_{\text{turb}}$  measured at each radius is comparable within an order of magnitude to the local cooling rate and, more importantly, scales linearly with it from radius to radius and between clusters. The answer, as demonstrated in Fig. 3, is ‘yes’. Here the gas cooling rate was evaluated directly from the measured gas density and temperature  $T$ :  $Q_{\text{cool}} = n_e n_i A_n(T)$ , where  $n_e$  and  $n_i$  are the number densities of electrons and ions, respectively, and  $A_n(T)$  is the normalized gas cooling function<sup>26</sup>. We see that, in all seven annuli in Perseus and all four in Virgo (which span the cluster cores in both cases),  $Q_{\text{turb}} \sim Q_{\text{cool}}$  over nearly three orders of magnitude in the value of either rate (Fig. 3 and Methods). We note that in Virgo and Perseus similar levels of  $Q_{\text{cool}}$  and  $Q_{\text{turb}}$  are attained at physically different distances from the cluster centres.



**Figure 2 | Measured amplitude of the one-component velocity  $V_{1,k}$  of gas motions versus wavenumber  $k$ .** The amplitude is shown for two different annuli in both Perseus (blue) and Virgo (red). The values are obtained from the power spectra of density fluctuations, derived from the X-ray images. The wavenumber  $k$  is related to the spatial scale  $l$  by  $k = 1/l$ . Solid-hatched regions show the range of scales where the measurements are least affected by systematic and statistical uncertainties (Methods). The width of each curve reflects the estimated  $1\sigma$  statistical and stochastic uncertainties. The dashed line is the Kolmogorov scaling  $k^{-1/3}$ .



**Figure 3 | Turbulent heating ( $Q_{\text{heat}}$ ) versus gas cooling ( $Q_{\text{cool}}$ ) rates in the Perseus and Virgo cores.** Each shaded region shows the heating and cooling rates estimated in a given annulus (top right, the innermost radius; bottom left, the outermost radius; Extended Data Fig. 3). The size of each rectangle reflects  $1\sigma$  statistical and stochastic uncertainties in heating, variations of the mean gas density and temperature across each annulus (affecting estimates of both cooling and heating) and the deviations of the measured spectral slope from the Kolmogorov law.

Although these results are encouraging, the uncertainties associated with the above analysis are, admittedly, large (Methods). It is difficult to prove unambiguously that we are dealing with a universal turbulent cascade, because other structures (for example edges of the bubbles, entrainment of hot bubble matter<sup>12</sup>, sound waves<sup>15,16</sup>, mergers and gas sloshing<sup>27</sup>) might also contribute to the observed density fluctuation spectra. Rather we argue simply that the cluster cores appear disturbed enough that if these disturbances are indeed due to turbulence, then its dissipation can reheat the gas. At the very least, one may treat the amplitudes calculated here (Fig. 2) as an upper limit on the turbulent velocities. One of the major tasks for future X-ray observatories, capable of measuring the line-of-sight gas velocities directly, will be to verify the accuracy of these velocity amplitudes.

With this caveat, the approximate balance of cooling and heating (Fig. 3) suggests that turbulent dissipation may be the key mechanism responsible for compensating for gas cooling losses and keeping cluster cores in an approximate steady state. Although AGN activity is not the only driver of gas motions (mergers or galaxy wakes can contribute as well<sup>28</sup>), it is plausible that AGNs have the dominant role in the central  $\sim 100$  kpc, where the cooling time is short. If this is true then our results support the self-regulated AGN feedback model<sup>10</sup>, in which unchecked cooling causes accelerated accretion onto the central black hole, which responds by increasing the mechanical output, presumably in the form of bubbles of relativistic plasma. The bubbles then rise buoyantly, exciting internal waves in particular<sup>11,29</sup>; the energy from them is converted into turbulence, which cascades to small scales and eventually dissipates, reheating the gas.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 2 July; accepted 29 August 2014.

Published online 26 October 2014.

1. Lea, S. M. The dynamics of the intergalactic medium in the vicinity of clusters of galaxies. *Astrophys. J.* **203**, 569–580 (1976).

2. Cowie, L. L. & Binney, J. Radiative regulation of gas flow within clusters of galaxies: a model for cluster X-ray sources. *Astrophys. J.* **215**, 723–732 (1977).
3. Fabian, A. C. & Nulsen, P. E. J. Subsonic accretion of cooling gas in clusters of galaxies. *Mon. Not. R. Astron. Soc.* **180**, 479–484 (1977).
4. Fabian, A. C. Cooling flows in clusters of galaxies. *Annu. Rev. Astron. Astrophys.* **32**, 277–318 (1994).
5. Peterson, J. R. & Fabian, A. C. X-ray spectroscopy of cooling clusters. *Phys. Rep.* **427**, 1–39 (2006).
6. Churazov, E., Forman, W., Jones, C. & Böhringer, H. Asymmetric, arc minute scale structures around NGC 1275. *Astron. Astrophys.* **256**, 788–794 (2000).
7. McNamara, B. R. & Nulsen, P. E. J. Heating hot atmospheres with active galactic nuclei. *Annu. Rev. Astron. Astrophys.* **45**, 117–175 (2007).
8. Fabian, A. C. Observational evidence of active galactic nuclei feedback. *Annu. Rev. Astron. Astrophys.* **50**, 455–489 (2012).
9. Birzan, L. et al. The duty cycle of radio-mode feedback in complete samples of clusters. *Mon. Not. R. Astron. Soc.* **427**, 3468–3488 (2012).
10. Churazov, E., Brüggemann, M., Kaiser, C. R., Böhringer, H. & Forman, W. Evolution of buoyant bubbles in M87. *Astrophys. J.* **554**, 261–273 (2001).
11. Omma, H., Binney, J., Bryan, G. & Slyz, A. Heating cooling flows with jets. *Mon. Not. R. Astron. Soc.* **348**, 1105–1119 (2004).
12. Hillel, S. & Soker, N. Heating cold clumps by jet-inflated bubbles in cooling flow clusters. Preprint at <http://arxiv.org/abs/1403.5137>.
13. Zhuravleva, I. et al. The relation between gas density and velocity power spectra in galaxy clusters: qualitative treatment and cosmological simulations. *Astrophys. J.* **788**, L13–L18 (2014).
14. Gaspari, M. et al. The relation between gas density and velocity power spectra in galaxy clusters: high-resolution hydrodynamic simulations and the role of conduction. *Astron. Astrophys.* **569**, A67–A82 (2014).
15. Fabian, A. C. et al. A very deep Chandra observation of the Perseus cluster: shocks, ripples and conduction. *Mon. Not. R. Astron. Soc.* **366**, 417–428 (2006).
16. Sternberg, A. & Soker, N. Sound waves excitation by jet-inflated bubbles in clusters of galaxies. *Mon. Not. R. Astron. Soc.* **395**, 228–233 (2009).
17. Brethouwer, G., Billant, P., Lindborg, E. & Chomaz, J.-M. Scaling analysis and simulation of strongly stratified turbulent flows. *J. Fluid Mech.* **585**, 343–368 (2007).
18. Arévalo, P., Churazov, E., Zhuravleva, I., Hernández-Monteagudo, C. & Revnivtsev, M. A Mexican hat with holes: calculating low-resolution power spectra from data with gaps. *Mon. Not. R. Astron. Soc.* **426**, 1793–1807 (2012).
19. Dennis, T. J. & Chandran, B. D. G. Turbulent heating of galaxy-cluster plasmas. *Astrophys. J.* **622**, 205–216 (2005).
20. Kolmogorov, A. N. The local structure of turbulence in incompressible viscous fluid for very large Reynolds' numbers. *Dokl. Akad. Nauk SSSR* **30**, 301–305 (1941).
21. Sreenivasan, K. R. On the universality of the Kolmogorov constant. *Phys. Fluids* **7**, 2778–2784 (1995).
22. Kaneda, Y., Ishihara, T., Yokokawa, M., Itakura, K. & Uno, A. Energy dissipation rate and energy spectrum in high resolution direct numerical simulations of turbulence in a periodic box. *Phys. Fluids* **15**, L21–L24 (2003).
23. Beresnyak, A. Spectral slope and Kolmogorov constant of MHD turbulence. *Phys. Rev. Lett.* **106**, 075001 (2011).
24. Schekochihin, A. A. et al. Astrophysical gyrokinetics: kinetic and fluid turbulent cascades in magnetized weakly collisional plasmas. *Astrophys. J. Suppl. Ser.* **182**, 310–377 (2009).
25. Churazov, E. et al. X-ray surface brightness and gas density fluctuations in the Coma cluster. *Mon. Not. R. Astron. Soc.* **421**, 1123–1135 (2012).
26. Sutherland, R. S. & Dopita, M. A. Cooling functions for low-density astrophysical plasmas. *Astrophys. J. Suppl. Ser.* **88**, 253–327 (1993).
27. Markevitch, M. & Vikhlinin, A. Shocks and cold fronts in galaxy clusters. *Phys. Rep.* **443**, 1–53 (2007).
28. Subramanian, K., Shukurov, A. & Haugen, N. E. L. Evolving turbulence and magnetic fields in galaxy clusters. *Mon. Not. R. Astron. Soc.* **366**, 1437–1454 (2006).
29. Balbus, S. A. & Soker, N. Resonant excitation of internal gravity waves in cluster cooling flows. *Astrophys. J.* **357**, 353–366 (1990).

**Acknowledgements** Support for this work was provided by the NASA through Chandra award number AR4-15013X issued by the Chandra X-ray Observatory Center, which is operated by the Smithsonian Astrophysical Observatory for and on behalf of the NASA under contract NAS8-03060. S.W.A. acknowledges support from the US Department of Energy under contract number DE-AC02-76SF00515. I.Z. and N.W. are partially supported from Suzaku grants NNX12AE05G and NNX13AI49G. P.A. acknowledges financial support from Fondecyt 1140304 and European Commission's Framework Programme 7, through the Marie Curie International Research Staff Exchange Scheme LACEGAL (PIRES-GA-2010-2692 64). E.C. and R.S. are partially supported by grant no. 14-22-00271 from the Russian Scientific Foundation.

**Author Contributions** I.Z.: data analysis, interpretation, manuscript preparation; E.C.: data analysis, interpretation, manuscript preparation; A.A.S.: interpretation, discussions, manuscript preparation; A.C.F.: principal investigator of the Perseus cluster observations, interpretation, manuscript review; S.W.A.: interpretation, discussions, manuscript review; W.R.F.: principal investigator of the M87 observations, interpretation, manuscript review; P.A., J.S.S., A.S., R.S., A.V., N.W.: interpretation, discussions and manuscript review.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to I.Z. ([zhur@stanford.edu](mailto:zhur@stanford.edu)).



## METHODS

**Data processing.** We use Chandra data ObsIDs 3209, 4289, 4946–4953, 6139, 6145, 6146, 11713–11716, 12025 and 12033–12037 for the Perseus cluster and ObsIDs 2707, 3717, 5826–5828, 6186, 7210–7212 and 11783 for the Virgo cluster to extract projected density fluctuation spectra in a set of radial annuli. The initial data processing has been done following the standard procedure<sup>30</sup>, applying the most recent calibration data. To obtain the thermodynamic properties of both clusters, the spectra are deprojected<sup>31</sup> and fitted in the 0.6–9 keV band, using the XSPEC<sup>32,33</sup> code and APEC plasma model based on ATOMDB version 2.0.1. The spectral modelling approximates the emission from each shell as a single-temperature plasma in collisional equilibrium and assumes a constant metal abundance of half the solar value<sup>34</sup>.

The X-ray mosaic image and its reduced counterpart for the Virgo cluster are shown in Extended Data Fig. 2. The 0.5–3.5 keV band was used because it contains the dominant fraction of the cluster signal and because of the weak temperature dependence of the gas emissivity in this band. The image of relative fluctuations is obtained by dividing the mosaic image by a spherically symmetric  $\beta$ -model of the mean surface brightness profile taking a core radius of  $0.34' = 1.7$  kpc and a slope of  $\beta = 0.39$ . Point sources have been excised from the images, using circles scaled according to the size of the combined point spread function. Extended Data Fig. 3 shows the set of annuli in Perseus and Virgo in which this analysis was performed.

**Mean profiles.** Deprojected radial profiles of the electron number density  $n_e$  and temperature  $T_e$  are shown in Extended Data Fig. 1 for both clusters. Note that the properties of the two clusters are very different. In particular, the density in Virgo is a factor of  $\sim 3$  (or more) lower than in Perseus at radii beyond  $\sim 10$  kpc. The temperature in Virgo is also lower, by a factor of 1.5–2 at  $r \approx 10$ –20 kpc. Yet  $Q_{\text{turb}} \sim Q_{\text{cool}}$  in both clusters, as shown in Fig. 3, suggesting a self-regulated mechanism such as, for example, the AGN feedback model<sup>35</sup>.

The mean mass density of the gas is  $\rho_0 = (n_e + n_i)\mu_p = \xi\mu_p n_e$ , where  $n_i = (\xi - 1)n_e$  is the ion number density and  $\mu_p$  is the proton mass. Consider a fully ionized plasma with an abundance of heavy elements 0.5 the solar value,  $\xi = 1.912$  and a mean particle weight of  $\mu = 0.61$ . The cooling time is defined as  $t_{\text{cool}} = \frac{3(n_e + n_i)k_B T}{2n_e n_i A_n(T)} = \frac{3\xi}{2\xi - 1} \frac{k_B T}{n_e A_n(T)}$ , where  $A_n(T)$  is the normalized cooling function<sup>26</sup> ( $\text{erg cm}^3 \text{s}^{-1}$ ),  $k_B$  is the Boltzmann constant and we assume identical ion and electron temperatures:  $T = T_e = T_i$ . The sound speed, treating the ICM as an ideal monatomic gas, is  $c_s = \sqrt{\frac{5}{3} \frac{k_B T}{\mu_p}}$ .

Both  $t_{\text{cool}}$  and  $c_s$  are plotted in Extended Data Fig. 1 as functions of radius. It is manifest that  $t_{\text{cool}}$  is shorter than the Hubble time in the central  $\sim 100$  kpc. Note that the cooling time is at least 7–20 times longer than the characteristic free-fall time  $t_{\text{ff}}$  in both clusters, defined in terms of the radius  $r$  and the gravitational acceleration  $g$  as  $t_{\text{ff}} = (2r/g)^{1/2}$ . Therefore, thermal instability is at most marginally important for the hot gas<sup>36</sup>.

The cooling time  $t_{\text{cool}}$  and the cooling rate  $Q_{\text{cool}}$  have been calculated using a cooling function  $A_n(T)$  for gas with solar metallicity. This is a conservative choice, because the dependence of the cooling function on metallicity is not strong and often can be neglected for typical ICM gas temperatures,  $\sim 2 \times 10^7$ – $10^8$  K. In addition, cluster-core metallicity measurements from X-ray spectra can be biased owing to the complexity of the spectral modelling of multi-temperature plasma. By accounting for radial metallicity variations in both clusters (based on the simplest one-temperature spectral model) and the consequent variation of the cooling function, the cooling rates shown in Fig. 3 and Extended Data Fig. 4 may be lower by a factor of 0.8 in Perseus and in the outermost annuli in Virgo, but higher by a factor of 2 in the innermost annuli in Virgo.

**Estimates of velocity and density required for heating-cooling balance.** It is useful to have a-priori estimates of the fluctuation amplitudes required to make a heating-cooling balance plausible. Equating  $Q_{\text{cool}} = n_e n_i A_n(T)$  and  $Q_{\text{turb}} = C_0 \rho_0 V_{1,k}^3 k$ , the characteristic Mach number of the turbulent motions at scale  $l = 1/k$  becomes

$$\text{Ma} = \sqrt{3} \frac{V_{1,k}}{c_s} = \sqrt{3} \left( \frac{\xi - 1}{\xi} \frac{A_n(T)}{\mu_p C_0} \right)^{1/3} n_e^{1/3} c_s^{-1} k^{-1/3} \\ \approx 0.15 \left( \frac{n_e}{10^{-2} \text{ cm}^{-3}} \right)^{1/3} \left( \frac{c_s}{1,000 \text{ km s}^{-1}} \right)^{-1} \left( \frac{l}{10 \text{ kpc}} \right)^{1/3}$$

Here we have referred all of the equilibrium quantities to their typical order-of-magnitude values and used the fact that the normalized cooling function  $A_n(T)$  is a weak function of the ICM temperature<sup>26</sup>, allowing us to adopt the mean value  $A_n \approx 2.5 \times 10^{-23} \text{ erg cm}^3 \text{s}^{-1}$  (for a gas with solar metallicity). Because bubbles have typical sizes of  $\sim 5$ –20 kpc (ref. 8), the value  $l \sim 10$  kpc is a reasonable order-of-magnitude estimate of the outer scale for the ICM turbulence driven by such bubbles in cluster cores. Thus, the dissipation of turbulence with relatively low

Mach numbers,  $\text{Ma} \sim 0.15$ , should be sufficient to balance the cooling of the gas in cores.

In view of the relationship  $\delta\rho_k/\rho_0 \approx \eta_1 V_{1,k}/c_s$  between the amplitudes of density and velocity fluctuations<sup>13</sup>, these Mach numbers correspond to  $\delta\rho/\rho_0 \sim 10\%$ . These are indeed typical values of density fluctuations we see in galaxy clusters.

**Trivial part of the correlation between heating and cooling.** Because the density explicitly enters the expressions for both the cooling rate and the turbulent heating rate, the linear correlation between these rates seen in Fig. 3 partly reflects the large range of mean densities at different radii (Extended Data Fig. 1). To show that the correlation is not due solely to this trivial part, we divide both  $Q_{\text{cool}}$  and  $Q_{\text{turb}}$  by the density  $\rho_0$  and thus obtain the cooling and heating rates per unit mass ( $\text{erg s}^{-1} \text{g}^{-1}$ ) (Extended Data Fig. 4). Although the range of values of both rates is now smaller, as expected, the correlation between them remains manifest.

**Systematic uncertainties in the measurement of density fluctuation amplitudes.** We start with the measurements of the surface brightness fluctuations based on broadband X-ray images<sup>25</sup> (I.Z. *et al.*, manuscript in preparation), using the  $\Delta$ -variance method<sup>18,37</sup>. The variance on scale  $l$  estimated using this method corresponds to a convolution of the original power spectrum with a broad filter. For a Kolmogorov-like power spectrum, the method can overestimate<sup>18</sup> the amplitude of fluctuations by  $\sim 25\%$ .

A more important source of uncertainty in the determination of the density power spectrum is the fact that dividing the cluster image into ‘perturbed’ and ‘unperturbed’ components is ambiguous, especially for a relatively steep perturbation spectrum like Kolmogorov’s, whose integrated power is dominated by the largest scales<sup>25,38</sup>. The  $\beta$ -model provides a reasonable description of the radial surface brightness profiles for Perseus and Virgo. It is therefore a sensible starting choice of unperturbed cluster model. Of course, more complicated models, for example projection of an ellipsoidal  $\beta$ -model or models with more sophisticated radial profiles, could be used as well. Adding more flexibility (more fitting parameters) to the model allows one to absorb more large-scale features of the image into the model surface brightness distribution. The net result of such improved fitting is that the measured power in the remaining perturbations will decrease on large scales, whereas the small-scale power will be less affected (provided the spectrum,  $E(k)$ , is not steeper than  $k^{-3}$ , which would correspond to the spectral tail of a smooth large-scale distribution; indeed, all our measured spectra are close to the Kolmogorov  $k^{-5/3}$  spectrum, which satisfies this constraint). This would cause the power spectrum to flatten at large scales. This model-dependent nature of the large scales is a feature of any division of the surface brightness variations into unperturbed and perturbed parts, including the case of the simplest  $\beta$ -model. This is why we expect that the estimates of the heating power based on turbulence measurements on small scales within the inertial range are probably more robust than estimates based on larger, outer scales. Our estimate of  $\varepsilon$  is thus not very susceptible to the choice of the underlying model of the mean surface brightness profile.

The reconstruction of the three-dimensional power spectrum of density fluctuations  $P_{3D}$  from the two-dimensional power spectrum of the surface brightness fluctuations  $P_{2D}$  is another source of uncertainty. The geometrical factor  $f_{2D \rightarrow 3D} = P_{2D}/P_{3D}$  depends on the radial profile of the surface brightness<sup>25</sup>. We use the mean value of  $f_{2D \rightarrow 3D}$  for each annulus and conservatively estimate the uncertainties by comparing it with the factors for the inner and outer radii of the same annulus. The maximal uncertainty does not exceed 20% except for the innermost region of Virgo.

The random nature of density fluctuations is another source of uncertainty. The spectra we calculate are based on squared amplitudes averaged over each annulus. Given a (expected) large degree of intermittency of density fluctuations and the limited spatial extent of the annuli, one might ask how representative and how statistically converged (that is, well sampled) such annular averages are. For example, analysing fluctuations in small patches within the  $3'$ – $4.5'$  (62–94 kpc) annulus in Perseus, we find  $\delta\rho_k/\rho_0$  at scales  $k^{-1} \approx 15$  kpc varying in a relatively broad range from 3% to 10%. This difficulty in relating the root mean squared turbulence level to what happens (and what is observed) in any given location is unavoidable because one always observes only a single realization of the fluctuating field. To achieve statistical convergence, we perform our averages in relatively wide annuli. The results we report are robust in the sense that choosing twice broader annuli does not change the conclusions.

A related problem is associated with the weighting scheme used to calculate the amplitude of the fluctuations within each annulus by averaging an image after applying a filter that selects perturbations with a given spatial scale. The exposure maps of the images are not uniform and the brightness of the cluster itself also varies substantially across each annulus. The optimal weighting scheme for the reduction of Poisson noise would require the weights to be  $w_1 \sim t_{\text{exp}} I_0$ , where  $t_{\text{exp}}$  is the exposure map and  $I_0$  is the global  $\beta$ -model profile of the surface brightness. This means that those parts of the cluster that have higher numbers of counts would have larger weights. We have experimented with two other choices of weights:  $w_2 \sim t_{\text{exp}}$



and  $w_3 = 1$ . These weights have larger statistical errors, but provide a more uniform scheme for evaluating the amplitudes of the surface brightness fluctuations across the image. For the analysis reported in this Letter, we used the uniform weight  $w_3 = 1$ . In most cases (except for the innermost regions of the two clusters), the uncertainty associated with the choice of the weights does not exceed 20%.

The vertical width ('error bars') of the spectra shown in Fig. 2 and Extended Data Fig. 4 reflects the  $1\sigma$  statistical uncertainty. The uncertainties discussed above slightly affect the shape of the spectra and may change the normalization by the factors estimated above (I.Z. *et al.*, manuscript in preparation). The dark-shaded regions of the spectra in Fig. 2 and Extended Data Fig. 4b show the wavenumber ranges over which we deem the spectra to be determined reliably—these ranges were used to determine the turbulent cascade rate  $\varepsilon$  in the manner described in the main text. The high- $k$  limits of these ranges are set by the 'statistical' uncertainty (Poisson noise) or by the point spread function distortions of the amplitude (in both cases the uncertainty is less than 20% in the 'reliable' range). At low  $k$ , we limit our reliable  $k$  ranges by the wavenumbers where the spectra start flattening. The shape of the spectra at these scales is most probably determined by the presence of several characteristic length scales (for example distance from the cluster centre and scale heights) and by the large-scale uncertainties inherent in the choice of the underlying model of the unperturbed cluster and in using finite-width annular averaging regions. This flattening disappears or shifts to smaller  $k$  if thicker annuli are used.

**Systematic uncertainties in the density–velocity amplitude conversion.** If the perturbations of the intracluster gas are small, one expects a linear relationship between the velocity  $V_{1,k}$  and density  $\delta\rho_k/\rho_0$  spectral amplitudes<sup>13</sup>,  $\frac{\delta\rho_k}{\rho_0} \approx \eta_1 \frac{V_{1,k}}{c_s}$ , with  $\eta_1 \sim 1$  set by gravity wave physics. This assumes that the injection scale of the turbulence is larger than or comparable to the Ozmidov scale<sup>39</sup>—the scale on which the turbulent eddy turnover timescale becomes shorter than the buoyancy (Brunt–Väisälä) timescale (that is, nonlinear advection becomes more important than the buoyancy response). Dimensionally, this scale is  $l_0 = N^{-3/2} \varepsilon^{1/2}$ , where  $N = c_s/H$  is the Brunt–Väisälä frequency ( $H$  is the hydrostatic equilibrium scale height—we have omitted numerical factors and ignored the distinction between entropy, pressure and temperature scale heights) and  $\varepsilon = Q_{\text{turb}}/\rho_0$  is the turbulent cascade rate. The relationship  $\eta_1 \sim 1$  is inherited from large scales on all scales  $l < l_0$ , where the density becomes a passive scalar<sup>13</sup>.

Assuming that radiative cooling is balanced by turbulent heating,  $Q_{\text{turb}} = Q_{\text{cool}}$ , it is possible to make an a priori estimate of  $l_0$  by letting  $\varepsilon = Q_{\text{cool}}/\rho_0$  and using the local mean thermodynamic properties of the ICM to calculate  $Q_{\text{cool}}$ ,  $\rho_0$  and  $N$ . We have done this for both clusters, for each of the annuli where we subsequently calculated  $Q_{\text{turb}}$  (Extended Data Fig. 4). In all cases,  $l_0$  is within the range of scales (in some cases, comparable to the largest scales) over which velocity amplitudes were measured and used to calculate  $Q_{\text{turb}}$ , and for which the conclusion that  $Q_{\text{turb}} \sim Q_{\text{cool}}$  was drawn. Therefore, our assumption of  $\eta_1 = 1$  is at least self-consistent.

This assumption is also restricted to the inertial range, that is, to scales larger than any dissipative cut-offs. It is interesting to compare the smallest scales that we are probing with the Kolmogorov (dissipative) scale  $l_K = \nu^{3/4}/\varepsilon^{1/4}$ , where  $\nu$  is the kinematic viscosity calculated for unmagnetized gas (which is approximately the same as the parallel viscosity for a magnetized plasma<sup>40</sup>). In all regions considered in this work, the Kolmogorov scale is much smaller than the smallest scale used by us for the determination of the cascade rate. In the regions shown in Extended Data Fig. 4,  $l_K \approx 0.5$  and  $2$  kpc ( $k_K \approx 2$  and  $0.5$  kpc<sup>−1</sup>) in the 1.5'–3' and 3'–4.5' annuli in Perseus, respectively. In the Virgo cluster,  $l_K \approx 0.3$  and  $0.8$  kpc ( $k_K \approx 3$  and  $1.3$  kpc<sup>−1</sup>) in the 2'–4' and 4'–6' annuli, respectively.

Cosmological simulations of galaxy clusters confirm that  $\eta_1 \approx 1$  with a scatter of 30% (ref. 13). Hydrodynamic simulations with controlled driving of turbulence also show that  $\eta_1 \approx 1$ , provided that thermal conduction is suppressed<sup>14</sup>. The 30% scatter in the value of  $\eta_1$  gives a factor of 0.3–2 uncertainty in the heating rate.

We conclude that the cumulative uncertainty in the estimated heating rate is a factor of  $\sim 3$ . Although this uncertainty is large, the approximate agreement between heating and cooling rates is an interesting result, reinforced by the fact that the two rates are not only numerically comparable to each other but are also linearly correlated. A more rigorous test will become possible with direct measurements of the velocity field by future X-ray observatories.

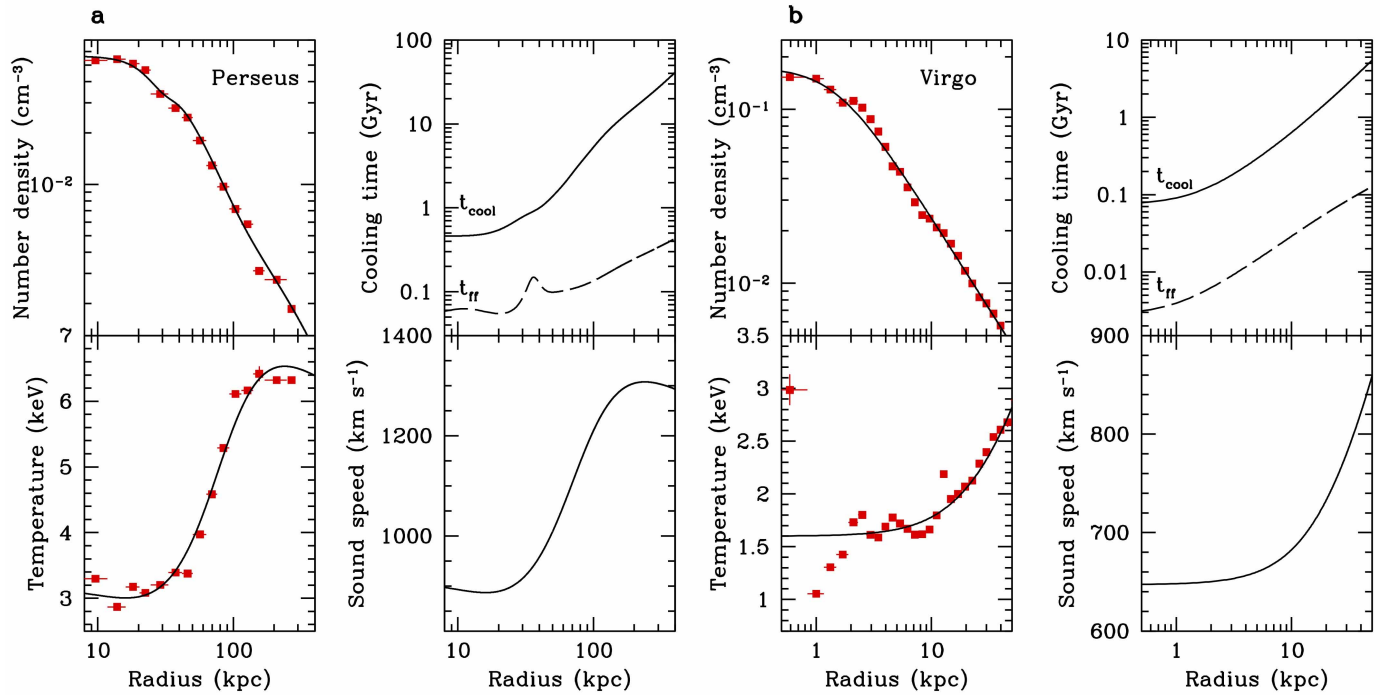
**Theoretical uncertainties: the nature of ripples and ICM heating theories.** Unsharp-masking of the Perseus image shows rough concentric rings, or 'ripples', in the surface brightness<sup>15</sup>. The observed morphology of these features, namely narrow in the radial direction and wide in the azimuthal direction, suggests two plausible possibilities: concentric sound waves<sup>15</sup> or stratified turbulence<sup>13,17</sup>. In the first case, the radial scale of the ripples should be determined by the time variability of the central AGN activity (factors include the intervals between outbursts, the excitation of multiple sound waves by vortices arising during each bubble inflation episode<sup>16</sup>, the distance from the centre and the ICM properties). In contrast, in the case of stratified turbulence, the radial scale  $\Delta r$  will be determined by the ratio of the characteristic

scale height  $H$  in the atmosphere and the velocity amplitude  $V$ :  $\Delta r \sim HV/c_s$ . Here we assume the second scenario and defer the detailed analysis of the nature of the substructure to a future publication.

Many other models of ICM heating, which could in principle offset radiative cooling in cluster cores, have been suggested. They differ widely in their presumed primary source of energy and in how this energy is channelled to the ICM. A brief and incomplete list of the broad classes into which these models fall is as follows: (1) source: thermal energy of the cluster gas; channelling mechanism: conductive heat flux to the core<sup>41,42</sup>; (2) source: cluster mergers; channelling mechanism: turbulence<sup>28,43</sup>; (3) source: galaxy motions; channelling mechanism: turbulence<sup>28,29,44,45</sup>; (4) source: central AGN; channelling mechanism: shocks and sound waves<sup>15,46</sup>, turbulent dissipation<sup>47,48</sup>, turbulent mixing<sup>49</sup>, cosmic rays<sup>50,51</sup>, radiative heating<sup>52,53</sup>, mixing of gas between ICM and the hot content of bubbles<sup>12</sup> and so on.

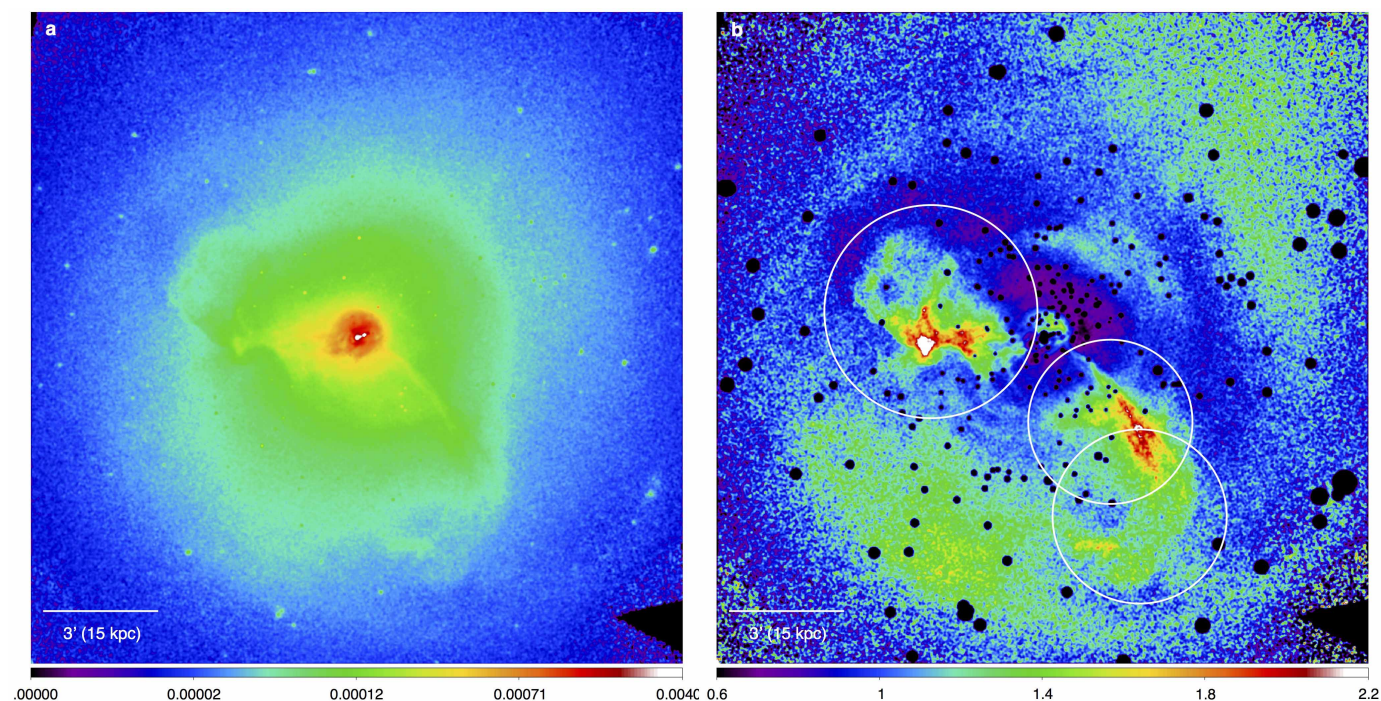
Given the multiplicity of possible scenarios, a detailed discussion and comparison of these models or even a complete list of references is beyond the scope of this Letter. We refer the reader the reviews in refs 7, 8 and references therein. The content of this Letter is focused on the energy channelling mechanism rather than the energy source. Note that, along with turbulent dissipation, turbulent heat conduction might also have a role in the cooling–heating balance. It can be shown, however, that in cluster cores, and assuming either stratified or isotropic turbulence, its contribution cannot be much larger than that of the turbulent heating (A.A.S. *et al.*, manuscript in preparation).

30. Vikhlinin, A. *et al.* Chandra temperature profiles for a sample of nearby relaxed galaxy clusters. *Astrophys. J.* **628**, 655–672 (2005).
31. Churazov, E., Forman, W., Jones, C. & Böhringer, H. XMM-Newton observations of the Perseus Cluster. I. The temperature and surface brightness structure. *Astrophys. J.* **590**, 225–237 (2003).
32. Foster, A. R., Ji, L., Smith, R. K. & Brickhouse, N. S. Updated atomic data and calculations for X-ray spectroscopy. *Astrophys. J.* **756**, 128–139 (2012).
33. Smith, R. K., Brickhouse, N. S., Liedahl, D. A. & Raymond, J. C. Collisional plasma models with APEC/APED: emission-line diagnostics of hydrogen-like and helium-like ions. *Astrophys. J.* **556**, L91–L95 (2001).
34. Anders, E. & Grevesse, N. Abundances of the elements: meteoritic and solar. *Geochim. Cosmochim. Acta* **53**, 197–214 (1989).
35. Churazov, E., Sunyaev, R., Forman, W. & Böhringer, H. Cooling flows as a calorimeter of active galactic nucleus mechanical power. *Mon. Not. R. Astron. Soc.* **332**, 729–734 (2002).
36. McCourt, M., Sharma, P., Quataert, E. & Parrish, I. J. Thermal instability in gravitationally stratified plasmas: implications for multiphase structure in clusters and galaxy haloes. *Mon. Not. R. Astron. Soc.* **419**, 3319–3337 (2012).
37. Ossenkopf, V., Krips, M. & Stutzki, J. Structure analysis of interstellar clouds. I. Improving the  $\Delta$ -variance method. *Astron. Astrophys.* **485**, 917–929 (2008).
38. Sanders, J. S. & Fabian, A. C. Deep Chandra and XMM-Newton X-ray observations of AWM 7 - I. Investigating X-ray surface brightness fluctuations. *Mon. Not. R. Astron. Soc.* **421**, 726–742 (2012).
39. Ozmidov, R. V. Length scales and dimensionless numbers in a stratified ocean. *Oceanology* **32**, 259–262 (1992).
40. Braginskii, S. I. Transport processes in a plasma. *Rev. Plasma Phys.* **1**, 205–310 (1965).
41. Zakamska, N. L. & Narayan, R. Models of galaxy clusters with thermal conduction. *Astrophys. J.* **582**, 162–169 (2003).
42. Cho, J. *et al.* Thermal conduction in magnetized turbulent gas. *Astrophys. J.* **589**, L77–L80 (2003).
43. Norman, M. L. & Bryan, G. L. in *The Radio Galaxy Messier 87* (eds Röser, H.-J. & Meisenheimer, K.) 106–115 (Springer, 1999).
44. Lufkin, E. A., Balbus, S. A. & Hawley, J. F. Nonlinear evolution of internal gravity waves in cluster cooling flows. *Astrophys. J.* **446**, 529–540 (1995).
45. Ruszkowski, M. & Oh, S. P. Galaxy motions, turbulence and conduction in clusters of galaxies. *Mon. Not. R. Astron. Soc.* **414**, 1493–1507 (2011).
46. Randall, S. W. *et al.* Shocks and cavities from multiple outbursts in the galaxy group NGC 5813: a window to active galactic nucleus feedback. *Astrophys. J.* **726**, 86–104 (2011).
47. Fujita, Y., Matsumoto, T. & Wada, K. Strong turbulence in the cool cores of galaxy clusters: can tsunamis solve the cooling flow problem? *Astrophys. J.* **612**, L9–L12 (2004).
48. Banerjee, N. & Sharma, P. Turbulence and cooling in galaxy cluster cores. *Mon. Not. R. Astron. Soc.* **443**, 687–697 (2014).
49. Kim, W.-T. & Narayan, R. Turbulent mixing in clusters of galaxies. *Astrophys. J.* **596**, L139–L142 (2003).
50. Chandran, B. D. & Dennis, T. J. Convective stability of galaxy-cluster plasmas. *Astrophys. J.* **642**, 140–151 (2006).
51. Pfrommer, C. Toward a comprehensive model for feedback by active galactic nuclei: new insights from M87 observations by LOFAR, Fermi, and H.E.S.S. *Astrophys. J.* **779**, 10–28 (2013).
52. Ciotti, L. & Ostriker, J. P. Cooling flows and quasars. II. Detailed models of feedback-modulated accretion flows. *Astrophys. J.* **551**, 131–152 (2001).
53. Nulsen, P. E. J. & Fabian, A. C. Fuelling quasars with hot gas. *Mon. Not. R. Astron. Soc.* **311**, 346–356 (2000).
54. Werner, N. *et al.* XMM-Newton high-resolution spectroscopy reveals the chemical evolution of M87. *Astron. Astrophys.* **459**, 353–360 (2006).



**Extended Data Figure 1 | Thermodynamic properties of the Perseus and Virgo clusters.** Radial profiles of the deprojected electron number density, the electron temperature, the cooling ( $t_{\text{cool}}$ ) and free-fall ( $t_{\text{ff}}$ ) times, and the sound speed. Red points: data with  $1\sigma$  error bars; black curves: data scatter approximations using smooth functions. The increased temperature scatter

in the central few kiloparsecs is associated with the presence of multi-temperature plasma in cool cores. A two-temperature fit of high-resolution XMM-Newton RGS spectra of the core of Virgo suggests an ambient temperature there of  $\sim 1.6$  keV (ref. 54). The smooth-function approximation we have chosen therefore approaches this value.

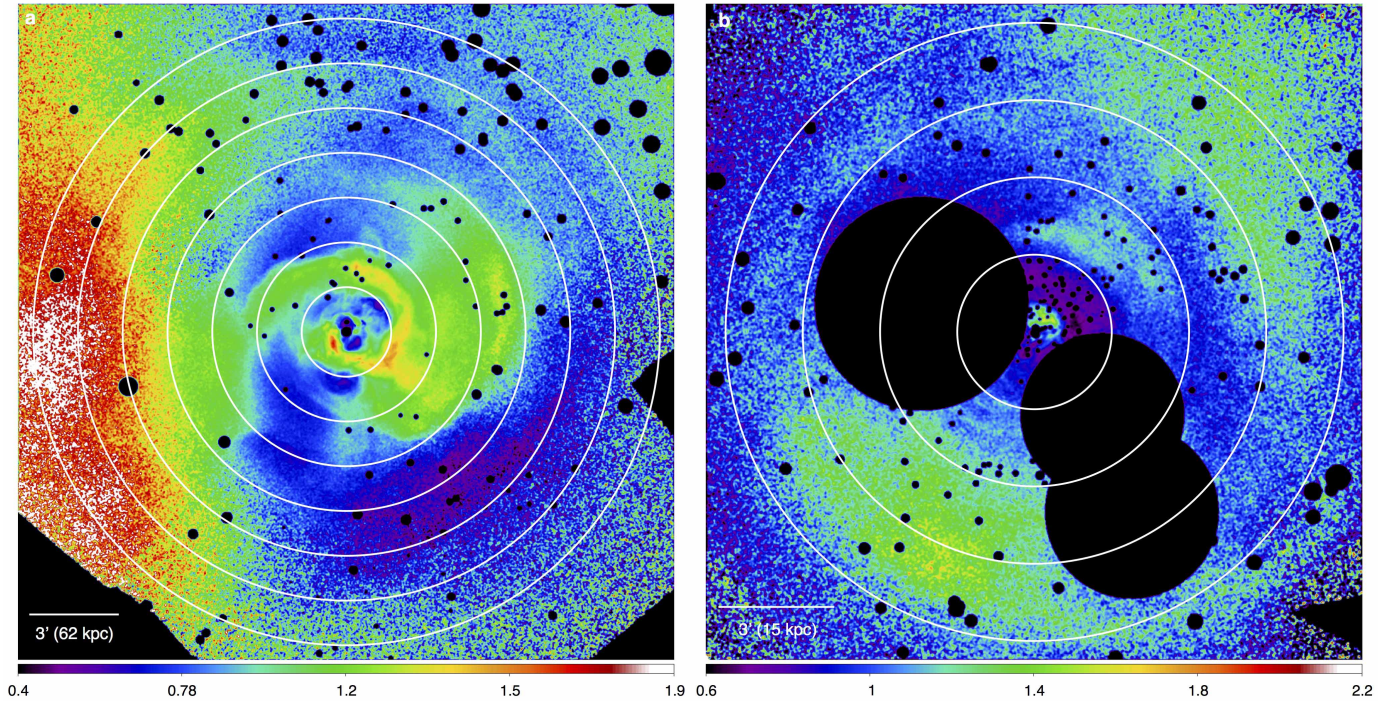


**Extended Data Figure 2 | X-ray image of the core of the Virgo cluster.**

**a**, X-ray surface brightness in units of counts per second per pixel in the 0.5–3.5 keV energy band. **b**, Relative surface brightness fluctuations. Both images are smoothed with a 3'' Gaussian. Black circles: excised point sources

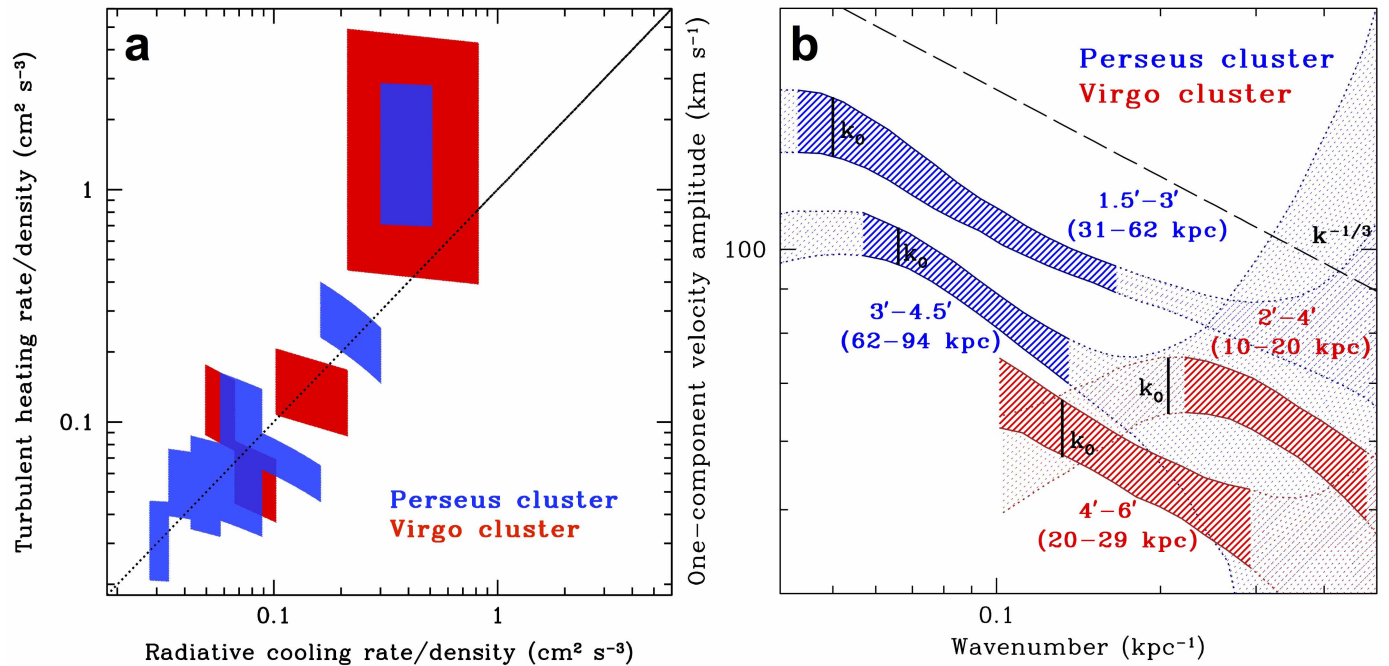
and central jet. White circles indicate 'arm-like' structures associated with the central AGN's activity, which have also been excised. We adopt 16.9 Mpc as the distance to the cluster, implying that an angular size of 1' corresponds to a length scale of 4.91 kpc.





**Extended Data Figure 3 | Set of the radial annuli used in the analysis of the Perseus and Virgo clusters.** The same as Fig. 1b and Extended Data Fig. 1b with white circles indicating the annuli used. The width of each annulus is

$1.5' \approx 31$  kpc in Perseus (a) and  $2' \approx 9.8$  kpc in Virgo (b). The outermost circles are  $10.5' \approx 218$  kpc and  $8' \approx 39$  kpc in Perseus and Virgo, respectively.



**Extended Data Figure 4 | Turbulent heating per unit density versus radiative cooling per unit density, and the Ozmidov scale in the Perseus and Virgo clusters.** **a**, The same as Fig. 3, but with the turbulent heating and cooling

rates divided by the mass density of gas in each annulus. **b**, The same as Fig. 2 with the Ozmidov scale  $l_O = 1/k_O = N^{-3/2} \varepsilon^{1/2}$  shown for each annulus (vertical black lines), estimated using  $\varepsilon = Q_{\text{cool}}/\rho_0$  (assuming that  $Q_{\text{turb}} = Q_{\text{cool}}$ ).

# Suppression of cooling by strong magnetic fields in white dwarf stars

G. Valyavin<sup>1</sup>, D. Shulyak<sup>2</sup>, G. A. Wade<sup>3</sup>, K. Antonyuk<sup>4</sup>, S. V. Zharikov<sup>5</sup>, G. A. Galazutdinov<sup>6,7</sup>, S. Plachinda<sup>4</sup>, S. Bagnulo<sup>8</sup>, L. Fox Machado<sup>5</sup>, M. Alvarez<sup>5</sup>, D. M. Clark<sup>5</sup>, J. M. Lopez<sup>5</sup>, D. Hiriart<sup>5</sup>, Inwoo Han<sup>9</sup>, Young-Beom Jeon<sup>9</sup>, C. Zurita<sup>10,11</sup>, R. Mujica<sup>12</sup>, T. Burlakova<sup>1</sup>, T. Szeifert<sup>13</sup> & A. Burenkov<sup>1</sup>

**Isolated cool white dwarf stars more often have strong magnetic fields than young, hotter white dwarfs<sup>1–4</sup>, which has been a puzzle because magnetic fields are expected to decay with time<sup>5,6</sup> but a cool surface suggests that the star is old. In addition, some white dwarfs with strong fields vary in brightness as they rotate<sup>7–10</sup>, which has been variously attributed to surface brightness inhomogeneities similar to sunspots<sup>8–12</sup>, chemical inhomogeneities<sup>13,14</sup> and other magneto-optical effects<sup>15–17</sup>. Here we describe optical observations of the brightness and magnetic field of the cool white dwarf WD 1953-011 taken over about eight years, and the results of an analysis of its surface temperature and magnetic field distribution. We find that the magnetic field suppresses atmospheric convection, leading to dark spots in the most magnetized areas. We also find that strong fields are sufficient to suppress convection over the entire surface in cool magnetic white dwarfs, which inhibits their cooling evolution relative to weakly magnetic and non-magnetic white dwarfs, making them appear younger than they truly are. This explains the long-standing mystery of why magnetic fields are more common amongst cool white dwarfs, and implies that the currently accepted ages of strongly magnetic white dwarfs are systematically too young.**

The complex magnetic field and photometric variability of the white dwarf star WD 1953-011<sup>18,19</sup> have been suggested<sup>8–12</sup> to result from a dark, magnetically generated spot on the star's surface. Using spectropolarimetric and photometric observations acquired by our group<sup>8,11,12</sup> from 2001 to 2009, and by other authors<sup>9,20</sup> we reconstructed the distribution of surface magnetic field intensity<sup>11</sup> and established<sup>12</sup> a physical relationship between an intense localized magnetic surface feature and a dark (cool) spot<sup>11,12</sup>. Relationships to sunspots were discussed, but remained speculative owing to the fundamental differences between atmospheres of white dwarfs and solar-type stars. Among these differences, the most remarkable are the high densities resulting from the compact nature of white dwarfs, their simple chemical compositions and the presence of extremely strong global magnetic fields in some stars<sup>2,3</sup>.

Here we establish a direct relationship between the magnetic field strength and temperature distribution on the surface of WD 1953-011. The detected cool/dark spot in this interpretation is merely the coolest region of a generally smooth temperature distribution associated with the global inhibition of convective energy transfer near the stellar surface. We modelled the observable photometric variation assuming a dependence of the local temperature  $T_{\text{loc}}$  at each point on the stellar surface and the local modulus of the magnetic field  $|H|_{\text{loc}}$  of the form  $T_{\text{loc}} \propto |H|_{\text{loc}}^{-\gamma}$  where  $\gamma$  is an arbitrary constant. The best agreement between the observed and predicted fluxes was found for  $\gamma = 0.059 \pm 0.004$ . As can be seen in Fig. 1, the fit is excellent, which favours the proposed link between the magnetic field intensity and the local temperature

of the photosphere. Figure 1 presents a sophisticated synthetic picture of a white dwarf obtained from observations, with details of its magnetic field and atmosphere.

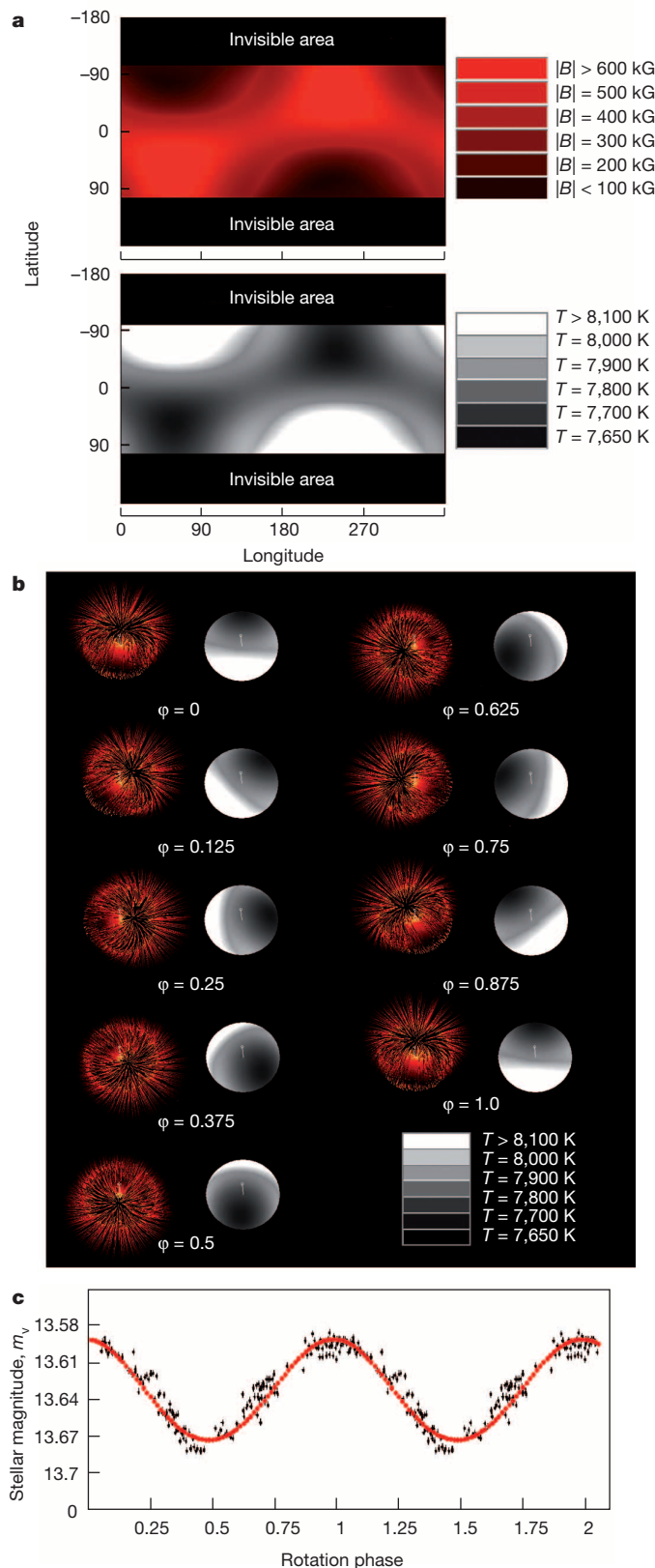
The link between the temperature and the magnetic field strength suggests that the most plausible explanation for the observed phenomenon is the effect known to operate in sunspots: the inhibition of convection by strong magnetic fields<sup>21</sup>. In the majority of white dwarfs with hydrogen atmospheres and temperatures below 12,000–14,000 K, convection transfers a significant fraction of the total energy flux. Modelling the atmosphere of WD 1953-011 using the LLmodels stellar atmosphere code<sup>22</sup> we estimate that convection transfers 70%–95% of the total flux in subphotospheric layers. This is a substantial fraction of the total flux and the suppression of convection should be capable of producing a temperature decrease comparable to that which causes sunspots. The power form of the adopted function  $T_{\text{loc}} \propto |H|_{\text{loc}}^{-\gamma}$  independently supports a suppression of convection similar to that occurring in sunspots<sup>21</sup>, and can be understood qualitatively—the stronger the field, the deeper the convective inhibition, and the cooler the upper photosphere. Using computations of the extent of the outer convection zone for hydrogen-rich stars<sup>23</sup> we also find that the magnetic field of WD 1953-011 is energetically sufficient to suppress convection over the entire stellar surface, and for some distance into the envelope. This, in turn, argues that the temperature–magnetic field relationship is valid everywhere over the surface.

Although the physical mechanism that produces cool, dark regions seems to be the same in sunspots and cool white dwarfs, there is a fundamental difference. According to theory<sup>24</sup>, generation of sunspots requires the presence of differential rotation and strong outer convection to transform a weak global magnetic field into strong-field flux tubes. These tubes are unstable and migrate over the stellar surface, and ultimately decay after tens to hundreds of days. In contrast to sunspots, the magnetic details and associated temperature/brightness distribution patterns in WD 1953-011 are stable and do not change over at least ten years<sup>12</sup>. The global control of surface convection by magnetic field is a fundamentally new observed effect having two important consequences.

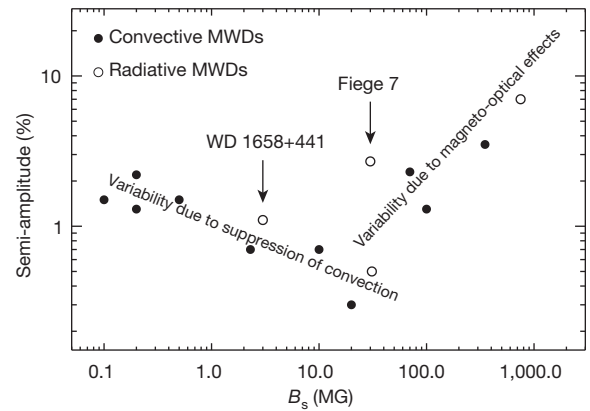
The first consequence is that the majority of convective magnetic white dwarfs (MWDs) should demonstrate stable, periodic photometric variability. By analysing known photometrically variable magnetic white dwarf stars<sup>10</sup>, we conclude that such variability is typical for the majority of those that are convective. The relationship between the observed<sup>10</sup> semi-amplitudes of the photometric variability of MWDs and their surface magnetic fields suggests the existence of two regimes or branches, as shown in Fig. 2. Fits to these branches using two different low-order polynomial functions demonstrate their statistically significant difference. We interpret the first, decreasing branch as due to the presence of

<sup>1</sup>Special Astrophysical Observatory of Russian Academy of Science, Nizhny Arkhiz, Zelenchukskiy region, Karachai-Cherkessian Republic 369167, Russia. <sup>2</sup>Institute for Astrophysics, Georg-August-University, Friedrich-Hund-Platz 1, D-37077, Göttingen, Germany. <sup>3</sup>Department of Physics, Royal Military College of Canada, PO Box 17000 Station Forces, Kingston, Ontario, K7K 7B4, Canada. <sup>4</sup>Crimean Astrophysical Observatory, Nauchny, Crimea 98409, Ukraine. <sup>5</sup>Instituto de Astronomía, Observatorio Astronómico Nacional San Pedro Martir (SPM), Universidad Nacional Autónoma de México, 22860 Ensenada, Baja California, México. <sup>6</sup>Instituto de Astronomía, Universidad Católica del Norte, Avenida Angamos 0610, Antofagasta, Chile. <sup>7</sup>Pulkovo Observatory, Pulkovskoe Shosse 65/1, Saint-Petersburg 196140, Russia. <sup>8</sup>Armagh Observatory, College Hill, Armagh BT61 9DG, UK. <sup>9</sup>Korea Astronomy and Space Science Institute 776, Daedeokdae-ro, Yuseong-gu, Daejeon 305-348, South Korea. <sup>10</sup>Instituto de Astrofísica de Canarias, 38200, La Laguna, Tenerife, Spain. <sup>11</sup>Departamento de Astrofísica, Universidad de La Laguna, 38206 La Laguna, Tenerife, Spain. <sup>12</sup>Instituto Nacional de Astrofísica, Óptica y Electrónica, Apartado Postal 51 y 216, 72000 Tonantzintla, Puebla, México. <sup>13</sup>European Southern Observatory, Alonso de Córdova 3107, Santiago, Chile.





**Figure 1 | Modelling results for WD 1953-011.** **a**, Tomographic portraits of the star's surface magnetic field (the field modulus, shown in red shades) and temperature distribution (greyscale) in planar projection. **b**, Tomographic portraits of the star's surface magnetic field (red) and temperature distribution (greyscale) at different rotation phases  $\varphi$  (spherical coordinates). **c**, Phase variation of brightness in stellar magnitudes (on the vertical axis) with rotation period  $P = 1.441788$  days of the V-band magnitude of WD 1953-011. Observations are shown as black symbols and their uncertainties<sup>12</sup>; the model is shown as red symbols.



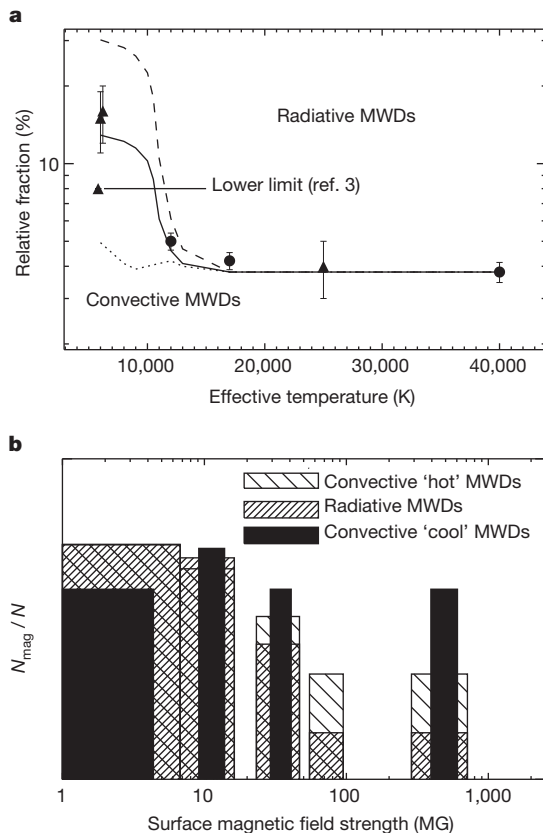
**Figure 2 | Dependence of amplitude of the photometric variability of MWDs on their surface magnetic field strength  $B_s$ .** MWDs form two sequences of variability: radiative and convective stars. The only exception is the very hot magnetic white dwarf WD 1658+441, the variability of which remains mysterious<sup>10</sup>. This young degenerate is still forming, and is located far before the crystallization stage; we cannot exclude the possibility that it is a binary or has an orbiting planet. The star Feige 7 is a unique MWD, the variability of which is due to magnetically separated surface abundances of helium and hydrogen<sup>13,14</sup>.

global suppression of convection. For a more detailed interpretation of Fig. 2 see the Methods.

The second consequence is that the cooling evolution of white dwarfs with the strongest magnetic fields is slowed. This surprising effect explains several mysterious statistical characteristics of MWDs. According to the white dwarf cooling theory<sup>25</sup> the characteristic cooling timescale  $\tau$  is defined as  $\tau = (L/M)^{-5/7}$ , where  $L$  and  $M$  are the star's luminosity and mass. As noted above, in white dwarfs convection typically transfers 70%–95% of the flux from subphotospheric layers to the surface. Therefore, strongly magnetic, convective white dwarfs with sufficiently inhibited convection should typically have much smaller luminosities, and therefore longer cooling timescales than their non-magnetic or weakly magnetic twins. Following published studies<sup>26,27</sup>, in the Methods we present theoretical details on convection inhibition and qualitatively model the process. We argue that convection is inhibited in all cool white dwarfs with magnetic fields from hundreds of kilogauss to tens of mega-gauss and higher. As a consequence their luminosities are decreased (especially for those with the strongest fields), and their cooling timescales are increased. (We note that hydrostatic equilibrium is always preserved because white dwarfs are cooling remnants without burning cores.)

The latter conclusion suggests that magnetic white dwarfs must exhibit characteristic features in their temperature–age distributions related to these phenomena. Indeed, statistical studies demonstrate that the observed<sup>1–3</sup> fraction of strongly magnetic stars among cool white dwarfs is higher than among the hot ones. Whether this is so has been historically debated<sup>2,3</sup>, but the question has recently been settled by the discoveries of hundreds of magnetic degenerate stars in the Sloan Digital Sky Survey<sup>4</sup>. In Fig. 3a we show the frequency versus effective temperature of magnetic white dwarfs from this survey, and from a volume-limited sample<sup>1–3</sup>. The increase in frequency for magnetic stars in the convective region is obvious. Having been firmly established, this effect is still not understood.

We suggest that the magnetic suppression of cooling provides a natural explanation of the frequency difference between cool (convective) and hot (convection-free) MWDs. At the same time, the frequency difference is not found among the weak-field stars with fields<sup>28,29</sup> of a few kilogauss (kG). For these degenerates, convection is not suppressed by the field and the cooling suppression does not occur. The basic consequence is that the slower cooling evolution of cool MWDs changes their distribution relative to hot ones, leading the majority of the strong-field stars to occur in the low-temperature regime. Using predictions

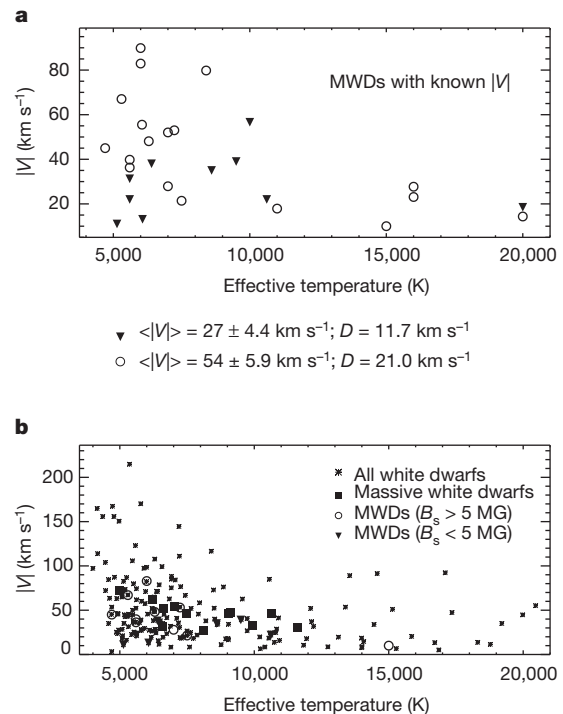


**Figure 3 | Some characteristic features in statistical distributions of MWDs.** **a**, Frequency dependence of MWDs on temperatures. Filled circles are frequencies from the most recent survey<sup>4</sup>. Triangles are frequencies from other studies<sup>1–3</sup>. Uncertainties are obtained by the Monte-Carlo method and taken from the literature<sup>1–3</sup>. The solid and dashed lines are computations obtained assuming inhibition of convection in only the photosphere (dotted line); in the upper subphotospheric layers and photosphere (solid line); and the entire convective zone (dashed line). **b**, Distributions of the relative fraction  $N_{\text{mag}}/N$  of MWDs sorted by their magnetic field strengths into three groups: convective, cool MWDs with temperatures below 10,000 K, convective, hot MWDs with temperatures from 10,000 K to 14,000 K, and radiative MWDs.  $N_{\text{mag}}$  and  $N$  are the observed numbers of magnetic white dwarfs and all white dwarfs in each of these groups.

of the LLmodels stellar atmosphere code<sup>22</sup>, we compute three idealized cases of changes in fractional incidence of magnetism due to convection inhibition among white dwarfs. The results are presented in Fig. 3a. Details of these computations can be found in the Methods. One of these predictions (the solid line) exhibits excellent agreement with observations, theoretically supporting our findings.

In fact, the increasing fraction of strongly magnetic white dwarfs with age can also be inferred from inspection of published samples<sup>4,28</sup> (Fig. 3b). Finally, the effect of slowing the evolution is directly seen in Fig. 4, where we plot observed<sup>3,30</sup> spatial motions of MWDs arranged by their surface field strengths into two groups. We find that the cool/convective stars show a statistically significant increase in their space velocities and velocity dispersions with field strength, indicative<sup>30</sup> of increasing age with increasing field strength. Unfortunately, a similar comparison of the MWD population with the population of non-MWDs is complicated in several ways, as explained and analysed in the Methods. From this analysis we conclude that the coolest white dwarfs with the strongest fields are the oldest white dwarfs in the Galaxy. This conclusion and possible alternative explanations of the statistical phenomena under consideration are also discussed in the Methods.

The arguments considered above support the existence of a magnetic suppression of cooling in strongly magnetic, isolated white dwarfs. This prompts a revision of our interpretation of the MWD cooling sequence



**Figure 4 | Absolute values of spatial velocities  $|V|$  of MWDs versus their surface temperatures.** **a**, The sample of all MWDs with known<sup>3,30</sup>  $|V|$ . **b**, Analysis of the mean absolute velocities  $\langle |V| \rangle$  and dispersions  $D$  of MWDs. The difference  $\Delta \langle |V| \rangle$  in the mean velocities between the group of cool MWDs with field strengths lower than 5 MG (filled triangles) and the group of MWDs with stronger fields (open circles) is  $\Delta \langle |V| \rangle = 27 \pm 7.36 \text{ km s}^{-1}$ . **b**, The sample of all known white dwarfs within 25 pc from Earth. See Methods for a detailed explanation of this plot.

that, in turn, may require tuning of our understanding of the evolution of the Galaxy and the Universe.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 10 May; accepted 2 September 2014.

Published online 19 October 2014.

- Valyavin, G. & Fabrika, S. White dwarfs magnetic fields evolution. *Astron. Soc. Pacif. Conf. Ser.* **169**, 206–209 (1999).
- Liebert, J., Bergeron, P. & Holberg, J. B. The true incidence of magnetism among field white dwarfs. *Astron. J.* **125**, 348–353 (2003).
- Sion, E. M. *et al.* The white dwarfs within 25 pc of the Sun: kinematics and spectroscopic subtypes. *Astron. J.* **147**, 129 (2014).
- Kepler, S. O. *et al.* Magnetic white dwarf stars in the Sloan Digital Sky Survey. *Mon. Not. R. Astron. Soc.* **429**, 2934–2944 (2013).
- Wendell, C. E., Van Horn, H. M. & Sargent, D. Magnetic field evolution in white dwarfs. *Astrophys. J.* **313**, 284–297 (1987).
- Muslimov, A. G., Van Horn, H. M. & Wood, M. A. Magnetic field evolution in white dwarfs: the Hall effect and complexity of the field. *Astrophys. J.* **442**, 758–767 (1995).
- Barstow, M. A. *et al.* RE J0317–853: the hottest known highly magnetic DA white dwarf. *Mon. Not. R. Astron. Soc.* **277**, 971–985 (1995).
- Wade, G. A. *et al.* The magnetic white dwarf WD 1953–011: migrating magnetic and brightness spots? *Astron. Soc. Pacif. Conf. Ser.* **307**, 569–572 (2003).
- Brinkworth, C. S. *et al.* Rotational period of WD 1953–011—a magnetic white dwarf with a star-spot. *Mon. Not. R. Astron. Soc.* **357**, 333–337 (2005).
- Brinkworth, C. S., Burleigh, M. R., Lawrie, K., Marsh, T. R. & Knigge, C. Measuring the rotational periods of isolated magnetic white dwarfs. *Astrophys. J.* **773**, 47 (2013).
- Valyavin, G. *et al.* The peculiar magnetic field morphology of the white dwarf WD 1953–011: evidence for a large-scale magnetic flux tube? *Astrophys. J.* **683**, 466–478 (2008).
- Valyavin, G. *et al.* A study of the photometric variability of the peculiar magnetic white dwarf WD 1953–011. *Astrophys. J.* **734**, 17 (2011).
- Liebert, J., Angel, J. R. P., Stockman, H. S., Spinrad, H. & Beaver, E. A. Feige 7—a hot, rotating magnetic white dwarf. *Astrophys. J.* **214**, 457–470 (1977).

14. Achilleos, N., Wickramasinghe, D. T., Liebert, J., Saffer, R. A. & Grauer, A. D. Exploring the peculiar magnetic field of Feige 7. *Astrophys. J.* **396**, 273–288 (1992).
15. Ferrario, L., Vennes, S., Wickramasinghe, D. T., Bailey, J. A. & Christian, D. J. EUVE J0317–855 A rapidly rotating, high-field magnetic white dwarf. *Mon. Not. R. Astron. Soc.* **292**, 205–217 (1997).
16. Martin, B. & Wickramasinghe, D. T. Cyclotron absorption in magnetic white dwarfs. *Mon. Not. R. Astron. Soc.* **189**, 69–77 (1979).
17. Wickramasinghe, D. T. & Martin, B. Magnetic blanketing in white dwarfs. *Mon. Not. R. Astron. Soc.* **223**, 323–340 (1986).
18. Schmidt, G. D. & Smith, P. S. A search for magnetic fields among DA white dwarfs. *Astrophys. J.* **448**, 305–314 (1995).
19. Koester, D., Dreizler, S., Weidemann, V. & Allard, N. F. Search for rotation in white dwarfs. *Astron. Astrophys.* **338**, 612–622 (1998).
20. Maxted, P. F. L., Ferrario, L., Marsh, T. L. & Wickramasinghe, D. T. W. D. 1953–011: a magnetic white dwarf with peculiar field structure. *Mon. Not. R. Astron. Soc.* **315**, L41–L44 (2000).
21. Solanki, S. K. Sunspots: an overview. *Astron. Astrophys. Rev.* **11**, 153–286 (2003).
22. Shulyak, D., Tsybal, V., Ryabchikova, T., Stütz, Ch. & Weiss, W. W. Line-by-line opacity stellar model atmospheres. *Astron. Astrophys.* **428**, 993–1000 (2004).
23. D'Antona, F., & Mazzitelli, I. White dwarf external layers. *Astron. Astrophys.* **74**, 161–171 (1979).
24. Parker, E. N. *Cosmical Magnetic Fields, their Origin and their Activity* 207–214 (Clarendon Press, 1979).
25. Shapiro, S. L. & Teukolsky, S. A. *Black Holes, White Dwarfs, and Neutron Stars: the Physics of Compact Objects* 82–105 (Wiley, 1983).
26. Cowling, T. G. Stellar structure—stars and stellar systems. *Comp. Astron. Astrophys.* **8**, 425–463 (1965).
27. Chandrasekhar, S. On the inhibition of convection by a magnetic field. *Phil. Mag.* **43**, 501–532 (1952).
28. Kawka, A., Vennes, S., Schmidt, G. D., Wickramasinghe, D. T. & Koch, R. Spectropolarimetric survey of hydrogen-rich white dwarf stars. *Astrophys. J.* **654**, 499–520 (2007).
29. Landstreet, J. D. *et al.* On the incidence of weak magnetic fields in DA white dwarfs. *Astron. Astrophys.* **545**, A30 (2012).
30. Anselowitz, T., Wasatonic, R., Matthews, K., Sion, E. M. & McCook, G. P. The parentage of magnetic white dwarfs: implications from their space motions. *Publ. Astron. Soc. Pacif.* **111**, 702–708 (1999).

**Acknowledgements** G.V. thanks J. Landstreet and S. Fabrika for discussions and practical help in interpreting the results. G.V. also thanks E. Kaisina for help in the preparation of the manuscript. G.V. and G.A.G. acknowledge the support of Chilean fund FONDECYT-regular (project 1120190). G.V. and D.H. acknowledge financial support from CONACyT, Mexico (grant 180817). G.V., T.B. and A.B. acknowledge financial support from the ministry of science and education of the Russian Federation (contracts 14.518.11.7070 and 16.518.11.7073). D.S. acknowledges financial support from CRC963 Astrophysical Flow Instabilities and Turbulence (project A16-A17). G.A.W. is supported by a Natural Sciences and Engineering Research Council (NSERC Canada) Discovery Grant. S.V.Z. acknowledges support from DGAPA/PAPIIT IN100614 and CONACYT 151858 projects. L.F.M. acknowledges financial support from the Universidad Nacional Autónoma de México under grant PAPIIT IN104612.

**Author Contributions** G.V., D.S. and G.A.W. analysed the main ideas, drew the basic conclusions presented in this study and wrote the text. G.V., D.S. and S.B. modelled the magnetic and atmospheric properties of WD 1953-011. All authors participated in the organization, conduct and reduction of spectral, spectropolarimetric and photometric observations of WD 1953-011 at different European, Russian, Ukrainian and Mexican observatories between 2002 and 2014.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to G.V. ([gvalyavin@sao.ru](mailto:gvalyavin@sao.ru)).



## METHODS

**Relationships in Fig. 2.** We believe that the two relationships (branches in Fig. 2) between the observed<sup>10</sup> semi-amplitudes of the photometric variability of MWDs and their surface magnetic field strengths reveal two different processes. In the first branch, variability decreases with field strength; we interpret this to be due to the presence of global suppression of convection. At the upper end of the first branch, for stars with magnetic fields between several hundred kilogauss and one megagauss, the photometric variability shows maximum amplitudes. Following this maximum the amplitudes begin to decrease, owing to a saturation of the effect with increasing field. Stars with stronger fields may become increasingly variable owing to changes in the character of the convection inhibition (see below) and/or due to other mechanisms such as opacity dichroism<sup>15</sup>, variable cyclotron absorption<sup>16</sup>, blanketing effect<sup>17</sup> and so on. All together, these mechanisms form the second, rising branch in Fig. 2. Two remarkable MWDs in Fig. 2 (WD 1658+441 and Feige 7) require a brief discussion. The ultra-massive, very hot star WD 1658+441 cannot be classified within this study—its variation has an alternative explanation<sup>10</sup>. The hot star Feige 7 is marked by an open circle in Fig. 2 despite the fact that this star is convective due to its helium-rich atmosphere<sup>13</sup>. Studies<sup>13,14</sup> have shown that this unique white dwarf exhibits a non-uniform surface helium abundance distribution produced by partially suppressed convection controlled by its very strong magnetic field, that is, the brightness variation of Feige 7 is also not caused by the mechanism we propose here.

**Details on modelling.** In this study, the observed brightness variation of WD 1953-011 was modelled numerically by testing an empirical dependence of the local surface temperatures  $T_{\text{loc}}$  on the local modulus of the magnetic field  $|H|_{\text{loc}}$  of the form  $T_{\text{loc}} = A|H|_{\text{loc}}^{-\gamma}$  where  $A$  and  $\gamma$  are arbitrary constants. The magnetic field intensities  $H_{\text{loc}}$  at each point of the stellar surface were taken from our previous studies<sup>11</sup>. The integration and minimization methods used to find optimal values of the arbitrary constants are standard, and are also described in the literature<sup>11</sup>.

To compute model atmospheres of white dwarfs we used the LLmodels stellar atmosphere code<sup>22</sup>. The code can treat strongly non-solar abundance patterns and therefore is applicable to white dwarf stars with pure hydrogen, helium or mixed compositions. The Stark-broadened profiles of hydrogen lines were computed using tables<sup>31</sup> based on the VCS theory<sup>32</sup>. The occupation probability formalism of Hummer and Mihalas<sup>33</sup> was used in computations of hydrogen level populations in dense plasma conditions. The convective energy transport was computed using the CM model of convection described by Canuto and Mazzitelli<sup>34,35</sup>. A commonly used treatment of convection based on mixing-length theory<sup>36</sup> is also provided in the code.

To calculate changes in fractional incidence of magnetism among white dwarfs due to convective inhibition we used the following relationships. In the first-guess approximation the fraction  $F(T)$  of MWDs in a given interval of temperatures is proportional to characteristic times<sup>25</sup> of the cooling evolution  $\tau(T) \sim (L(T))^{-5/7}$ . Computing white dwarf atmospheres using the atmosphere model code described above we estimate relative changes in luminosities  $L(T)$  taking convective energy transport into account and with inhibited convection  $L_i(T)$  considering different possible cases of inhibition. The resultant fractions  $F_i(T)$  of MWDs with inhibited convections were then recalculated relative to the observed fractions among radiative MWDs:

$$F_i(T) = F_r(L_i(T)/L(T))^{-5/7}$$

where  $F_r$  is the observed fraction of radiative MWDs (about 4%; ref. 4). Three cases of inhibition were considered (see Fig. 3a): when convection is fully inhibited in a MWD's atmosphere (above a Rosseland optical depth,  $\text{Tauross} = 1$ , short dashed line); when convection is inhibited at  $\text{Tauross} = 10$  (solid line); and when convection is fully inhibited (long dashed line). The case when convection is inhibited directly under the photosphere (the solid line) reproduces the observations well. This result characterizes a 'mean' case of convective inhibition resulting from an averaged value of magnetic field intensities of all known MWDs. This result also suggests that all cool MWDs typically increase their cooling times by a factor of two to three. In reality, however, MWDs with different magnetic field intensities are expected to reveal different efficiencies of the magnetic inhibition effect: the stars with weakest fields should be less affected than stars with stronger fields (see also below). This requires us to provide some additional explanation by qualitative consideration of the convective inhibition mechanisms and their characteristic relationship with magnetic field strength.

**Convective inhibition mechanisms.** According to studies<sup>26</sup>, magnetic fields profoundly affect convective energy transfer when the magnetic pressure/energy is comparable to the material pressure/energy in some layers of a star's convective zone. In these layers convective motions tend to stochastically twist magnetic field lines, causing the field to strongly resist the convective motions independently of the field orientation and proportionally to the square of the field modulus<sup>26</sup>. WD 1953-011 is a comparatively weak-field MWD, a good example of a star for which the

equality of magnetic and gas pressures occurs not very far beneath the photosphere. This is similar to what is observed in sunspots. Using numerical computations of external convective layers<sup>23</sup> we establish that this example represents MWDs with magnetic fields from a few hundred kilogauss to a few megagauss, for temperatures cooler than 8,000 K (if hydrogen-rich like WD 1953-011), or cooler than 16,000 K (if helium-rich). We define this group as weak-field MWDs with typical field strengths  $< 5$  MG.

For larger fields (tens of megagauss or more, which are common among the strongly magnetic MWDs) the similarity of magnetic and gas pressures is reached in very deep layers which correspond to the lower borders of the convective zone in cool hydrogen-rich MWDs with temperatures lower than 6,000 K, or for helium-rich MWDs with temperatures lower than 14,000 K. Above these layers the magnetic energy is much larger than the gas energy. In this case residual convection is present, but is restricted by the magnetic field to occur strictly along magnetic field lines<sup>13,14,26</sup>. We briefly consider this interesting case in more detail.

Because convective motions are redirected by magnetic fields along the magnetic lines, the efficiency of convective energy transport depends also on the orientation of the field. In those parts of the star's surface where the field is essentially horizontal (perpendicular to the gravity vector), vertical convection is reduced by the Lorentz force acting across magnetic field lines<sup>13,14</sup>. In regions corresponding to the poles of a dipole, where the field and gravity vectors are oriented mainly in the same direction, convection is inhibited through reduction of the convective instability, with a quadratic dependence on the field<sup>27</sup>. As a result<sup>27</sup>, convective cells are strongly elongated along magnetic field lines and the associated energy transport is reduced depending on the field strength. That is, we may also speak about global inhibition of convective energy transport with some efficiency that monotonously and asymptotically grows with field strength. However, the character of the expected photometric variability due to convective inhibition in strong-field MWDs may be different from the case of the weak-field star WD 1953-011. There are also a few intermediate cases of MWDs with comparatively thin convective zones (like Feige 7). For such stars inhibition of convective energy transport is expected to be inconsequential in the context of inhibition of cooling.

In general, accurate values of convective inhibition efficiency can be obtained only by three-dimensional magnetohydrodynamical calculations that are outside the scope of this study. Instead, the effectiveness of the influence of magnetic fields on inhibition of cooling and its dependence on an MWD's surface field can also be tested using observations, as illustrated in Fig. 4 and related explanations in the text. As can be seen, the group of cool weak-field MWDs with fields less than 5 MG exhibits significantly lower spatial velocities and velocity dispersions than the group of cool, strong-field stars. This argues that the cool, weak-field MWDs are younger than the strong-field MWDs of the same temperatures. It is also interesting to note that the inhibition of convection can be very effective in weak-field MWDs (for example, in WD 1953-011), but the effect works close to the surface in comparatively thin zones that makes the effect of magnetic inhibition of cooling less effective than in the strong-field MWDs ('thin inhibition'). In contrast, the strongest fields of MWDs entirely control the convective zone that, especially in cool stars, creates a 'thick inhibition' that slows their cooling evolution substantially, as shown in Fig. 4.

**Statistical analysis for Fig. 4b.** Finally, we discuss the kinematical properties of populations of MWDs and non-MWDs (which are probably white dwarfs with very weak fields below the detection limit). This comparison could additionally provide a test for the effect of magnetic suppression of cooling in strong-field MWDs. In general, this is not a trivial question because in contrast to the more or less homogeneous sample of MWDs originating mainly from intermediate-mass magnetic A/B stars<sup>30,37</sup>, non-MWDs are produced via a variety of different evolutionary channels<sup>30,38,39</sup>. Many of these channels may not produce MWDs. Among these channels the most populated is probably low/intermediate-mass white dwarfs originating from non-magnetic convective main-sequence stars of spectral classes later than A/B. These white dwarfs have essentially different space velocities distributed over a much larger range (Fig. 4b). Following the conventional point of view that MWDs are the remnants of intermediate-mass magnetic A/B stars<sup>30,37</sup> we may assume that the most massive non-MWDs of the same masses as MWDs originated from non-magnetic A/B stars. A comparison of the kinematical properties of MWDs and non-MWDs of the same masses could give an answer to the question of whether the magnetic population is older than the non-magnetic one. In Fig. 4b we plot those non-MWDs with masses exceeding 0.8 solar masses, which are masses typical of MWDs<sup>2</sup> (filled squares in Fig. 4). The mass estimates are taken from the most recent survey of white dwarfs with known trigonometric parallaxes<sup>40</sup>. As can be seen, these white dwarfs are located within the population of MWDs. Unfortunately, the sample of white dwarfs with known masses is very limited, complicating our analysis. Moreover, some of these white dwarfs might be from other evolutionary channels. In addition, the masses we used have significant uncertainties, typically about 0.15 solar masses at the  $3\sigma$  level, and may have some computational

biases owing to model simplifications. However, despite these problems, analysis of this population at least does not contradict our hypothesis that the strongly magnetic MWDs are the oldest white dwarfs of the same temperatures. The massive non-MWDs marked by squares in Fig. 4 demonstrate the characteristic velocity  $\langle |V| \rangle = 45.6 \pm 3.6 \text{ km s}^{-1}$  and a dispersion  $D = 13.0 \text{ km s}^{-1}$ . Comparison of these values with the corresponding values for the population of the strongly magnetic MWDs ( $\langle |V| \rangle = 54.6 \pm 5.9 \text{ km s}^{-1}$ ;  $D = 21.0 \text{ km s}^{-1}$ , see Fig. 4) additionally supports (especially comparison of the dispersions  $D$ ) our arguments that the strongly magnetic MWDs are older than the non-MWDs of the same masses and temperatures. This, in turn, supports our main finding, according to which cooling of strongly magnetic white dwarfs is suppressed by their magnetic fields. On the other hand, this analysis serendipitously provides the astronomical community with independent support of the idea<sup>37</sup> that MWDs are descended from intermediate-mass magnetic A/B stars.

Finally, we briefly comment on some alternative ideas related to the problems discussed in this paper. According to the relationship connecting cooling times, masses and luminosities (see above), it may be that the surface magnetic-field strengths of white dwarfs may depend on their masses (the larger the mass of a MWD, the smaller its radius and therefore the stronger its surface field). We may therefore expect an alternative relationship connecting masses, field strengths and ages. However, examination of this problem by other authors<sup>2</sup> as well as our own examination of MWDs plotted in Fig. 4 reveals no systematic differences in masses and gravities of groups of stars with different magnetic field strengths.

Other explanations for the statistical features in the distribution of MWDs illustrated in Fig. 3 include an evolutionary increase of surface magnetic fields of MWDs

due to an unknown mechanism of magnetic field generation<sup>1</sup>, or an evolution-induced diffusion of strong field from the subphotospheric envelope into the photosphere<sup>1,3</sup>. The existence of these mechanisms, however, has not been confirmed theoretically<sup>5,6</sup> or observationally in later studies<sup>28,29</sup>.

31. Lemke, M. Extended VCS Stark broadening tables for hydrogen–Lyman to Brackett series. *Astron. Astrophys.* **122** (Suppl.), 285–292 (1997).
32. Vidal, C. R., Cooper, J. & Smith, E. W. Hydrogen Stark-broadening tables. *Astrophys. J.* **25** (Suppl.), 37–135 (1973).
33. Hummer, D. G., & Mihalas, D. The equation of state for stellar envelopes. I—an occupation probability formalism for the truncation of internal partition functions. *Astrophys. J.* **331**, 794–814 (1988).
34. Canuto, V. M. & Mazzitelli, I. Stellar turbulent convection—a new model and applications. *Astrophys. J.* **370**, 295–311 (1991).
35. Canuto, V. M. & Mazzitelli, I. Further improvements of a new model for turbulent convection in stars. *Astrophys. J.* **389**, 724–730 (1992).
36. Böhm-Vitense, E. Über die Wasserstoffkonvektionszone in Sternen verschiedener Effektivtemperaturen und Leuchtkräfte. [in German] *Z. Astrophys.* **46**, 108–143 (1958).
37. Angel, J. R. P., Borra, E. F. & Landstreet, J. D. The magnetic fields of white dwarfs. *Astrophys. J.* **45** (Suppl.), 457–474 (1981).
38. Sion, E. M., Fritz, M. L., McMullin, J. P. & Lallo, M. D. Kinematical tests of white dwarf formation channels and evolution. *Astron. J.* **96**, 251–274 (1988).
39. Sion, E. M. & Liebert, J. The space motions and luminosity function of white dwarf. *Astrophys. J.* **213**, 468–478 (1977).
40. Bergeron, P., Legget, S. K. & Ruiz, M. T. Photometric and spectroscopic analysis of cool white dwarfs with trigonometric parallax measurements. *Astrophys. J.* **133** (Suppl.), 413–449 (2001).

# High-efficiency acceleration of an electron beam in a plasma wakefield accelerator

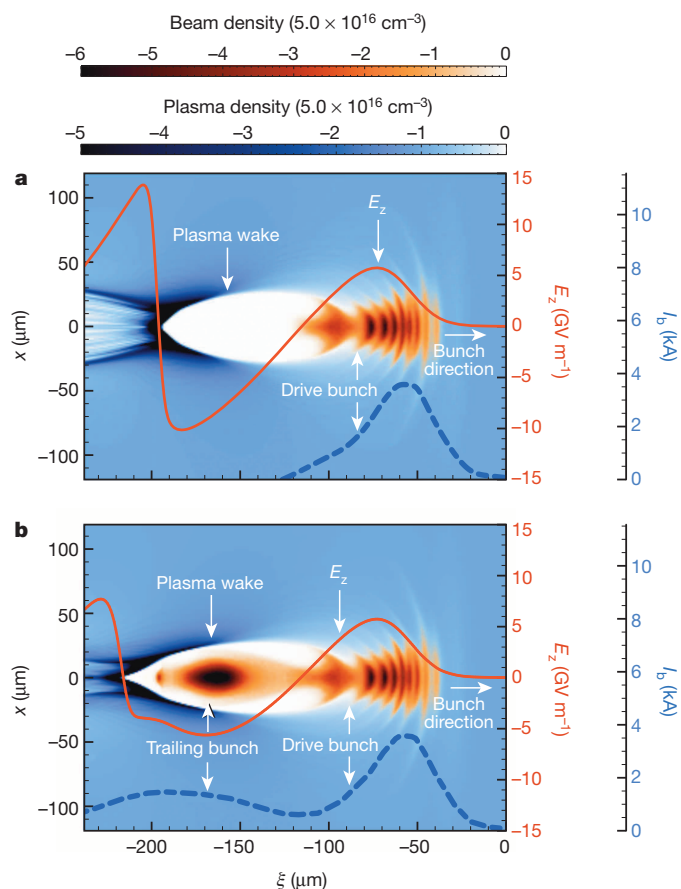
M. Litos<sup>1</sup>, E. Adli<sup>1,2</sup>, W. An<sup>3</sup>, C. I. Clarke<sup>1</sup>, C. E. Clayton<sup>4</sup>, S. Corde<sup>1</sup>, J. P. Delahaye<sup>1</sup>, R. J. England<sup>1</sup>, A. S. Fisher<sup>1</sup>, J. Frederico<sup>1</sup>, S. Gessner<sup>1</sup>, S. Z. Green<sup>1</sup>, M. J. Hogan<sup>1</sup>, C. Joshi<sup>4</sup>, W. Lu<sup>5</sup>, K. A. Marsh<sup>4</sup>, W. B. Mori<sup>3</sup>, P. Muggli<sup>6</sup>, N. Vafaei-Najafabadi<sup>4</sup>, D. Walz<sup>1</sup>, G. White<sup>1</sup>, Z. Wu<sup>1</sup>, V. Yakimenko<sup>1</sup> & G. Yocky<sup>1</sup>

High-efficiency acceleration of charged particle beams at high gradients of energy gain per unit length is necessary to achieve an affordable and compact high-energy collider. The plasma wakefield accelerator is one concept<sup>1–3</sup> being developed for this purpose. In plasma wakefield acceleration, a charge-density wake with high accelerating fields is driven by the passage of an ultra-relativistic bunch of charged particles (the drive bunch) through a plasma<sup>4–6</sup>. If a second bunch of relativistic electrons (the trailing bunch) with sufficient charge follows in the wake of the drive bunch at an appropriate distance, it can be efficiently accelerated to high energy. Previous experiments using just a single 42-gigaelectronvolt drive bunch have accelerated electrons with a continuous energy spectrum and a maximum energy of up to 85 gigaelectronvolts from the tail of the same bunch in less than a metre of plasma<sup>7</sup>. However, the total charge of these accelerated electrons was insufficient to extract a substantial amount of energy from the wake. Here we report high-efficiency acceleration of a discrete trailing bunch of electrons that contains sufficient charge to extract a substantial amount of energy from the high-gradient, nonlinear plasma wakefield accelerator. Specifically, we show the acceleration of about 74 picocoulombs of charge contained in the core of the trailing bunch in an accelerating gradient of about 4.4 gigavolts per metre. These core particles gain about 1.6 gigaelectronvolts of energy per particle, with a final energy spread as low as 0.7 per cent (2.0 per cent on average), and an energy-transfer efficiency from the wake to the bunch that can exceed 30 per cent (17.7 per cent on average). This acceleration of a distinct bunch of electrons containing a substantial charge and having a small energy spread with both a high accelerating gradient and a high energy-transfer efficiency represents a milestone in the development of plasma wakefield acceleration into a compact and affordable accelerator technology.

The experiment reported here is carried out in the three-dimensional, nonlinear regime of plasma wakefield acceleration, also known as the blow-out regime<sup>8</sup>. In this regime, a tightly focused and short ultra-relativistic electron bunch with a density greater than the plasma density propagates through a long column of plasma.  $\sigma_r \lesssim c/\omega_p$  and  $\sigma_z \lesssim \pi c/\omega_p$  are the root-mean-square (r.m.s.) transverse and longitudinal sizes of the beam, respectively, with  $\omega_p$  the plasma frequency. The transverse electric field of this drive bunch expels all of the plasma electrons within a radius of about 30  $\mu\text{m}$ , as shown in the three-dimensional particle-in-cell (QuickPIC<sup>9,10</sup>) simulation depicted in Fig. 1a. The Coulomb field of the stationary ions pulls the expelled plasma electrons back towards the central axis, which begins the wake oscillation, producing periodic ion cavities in the plasma. This wake structure follows the beam trajectory with a phase velocity matched to the drive bunch, at nearly the speed of light.

In the simulation, the input plasma and beam parameters are similar to those measured in the experiment with a simple scaling of the total beam charge (see Methods). The on-axis longitudinal electric field  $E_z$  of

the wake, also depicted in Fig. 1a, shows that the bulk of the drive bunch is located in a region of positive (forward-directed) electric field, and thus loses energy. If the electrons in the rear of such a drive bunch were to extend into the negative region of the electric field, they would gain energy from the wake. If there were not enough charge in the long tail of electrons to have a non-negligible impact on the profile of the steep



**Figure 1 | Three-dimensional particle-in-cell simulation of beam-driven plasma wakefield interaction.** **a**, A slice through the centre of an unloaded plasma wake, where  $x$  is the dimension transverse to the motion, and  $\xi = z - ct$  is the dimension parallel to the motion,  $E_z$  is the on-axis longitudinal electric field (red solid line) and  $I_b$  is the current of the input beam (blue dotted line). **b**, A plasma wake generated by the same drive bunch as in **a** when loaded by a trailing bunch. The plasma electron density is represented in blue, while the beam density is represented in red. The ion density (not shown) is uniform. The particle-in-cell code QuickPIC<sup>9,10</sup> was used to generate this simulation of the beam-plasma interaction.

<sup>1</sup>SLAC National Accelerator Laboratory, Menlo Park, California 94025, USA. <sup>2</sup>Department of Physics, University of Oslo, 0316 Oslo, Norway. <sup>3</sup>Department of Physics and Astronomy, University of California Los Angeles, Los Angeles, California 90095, USA. <sup>4</sup>Department of Electrical Engineering, University of California Los Angeles, Los Angeles, California 90095, USA. <sup>5</sup>Department of Engineering Physics, Tsinghua University, Beijing 100084, China. <sup>6</sup>Max Planck Institute for Physics, Munich 80805, Germany.



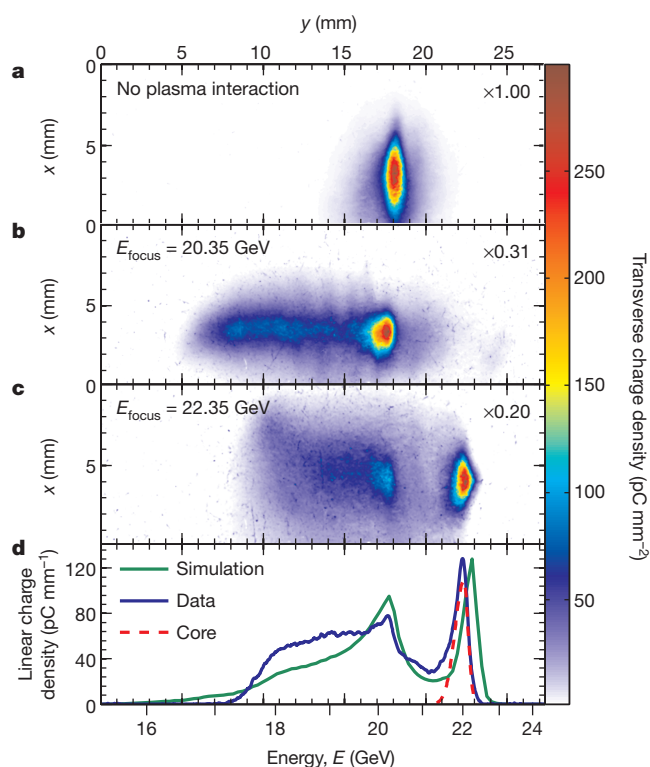
wakefield  $E_z$ , they would end up with a large final energy spread and leave behind a substantial amount of unused energy in the wake, as was the case in earlier experiments<sup>11–13</sup>.

The key to achieving a high energy-transfer efficiency with a narrow final energy spread is in placing an appropriately shaped trailing bunch at precisely the correct distance behind the drive bunch so as to flatten the wakefield  $E_z$  and provide for a uniform energy gain throughout the trailing bunch<sup>14,15</sup>. Theory and simulations have shown that a trailing bunch with a Gaussian distribution and the right charge content can lead to a nearly flattened wakefield<sup>8,15</sup>. Figure 1b shows an approximately double-Gaussian (in  $z$ ) input-charge distribution for both the drive and trailing bunches. The wakefield is noticeably altered by the addition of the trailing bunch charge, that is, the beam has loaded the wake. Note that the initial extent of the trailing bunch extends past the first ion cavity period (the region void of electrons), or ‘bucket’ of the wake. This leads to loss of charge because some of the electrons in the trailing bunch are strongly defocused by the returning plasma electrons at the back of the first bucket.

To reach accelerating gradients on the scale of more than a gigaelectronvolt per metre, the plasma medium must have a high electron density, which results in a correspondingly short plasma wavelength. This necessitates a very small separation between the drive and trailing bunches, and such a configuration is non-trivial to produce with ultra-relativistic electron beams<sup>16</sup>. The typical scale of a plasma wakefield acceleration wake structure at a plasma density of  $5 \times 10^{16} \text{ cm}^{-3}$  is around 200  $\mu\text{m}$ , as can be seen in Fig. 1, and therefore the separation between the charge-density peaks of the two bunches must be less than this value. The Facility for Advanced Accelerator Experimental Tests (FACET) at the SLAC National Accelerator Laboratory, where the experiment was carried out, was designed specifically to produce such high-current drive and trailing bunches with an appropriately small separation.

In the experiment, the SLAC linear accelerator provides a 20.35-GeV electron beam to the FACET experimental area, where it is then manipulated to form a two-bunch structure (see Methods) just before entering the plasma source, which is a 36-cm-long, laser-ionized column of lithium vapour with a density of  $5 \times 10^{16} \text{ cm}^{-3}$  contained inside a heat pipe oven<sup>17,18</sup> (see Methods). The final charge delivered to the plasma totals 1.80 nC, with 1.02 nC contained in the drive bunch ( $\sigma_z = 25 \mu\text{m}$ ,  $I_{\text{peak}} = 4.9 \text{ kA}$ ), and 780 pC contained in the broader trailing bunch ( $\sigma_z = 47 \mu\text{m}$ ,  $I_{\text{peak}} = 2.0 \text{ kA}$ ). Not all of this charge is coupled into the plasma wake, however, owing to the oversized width and length of the trailing bunch. After exiting the plasma, the spectrum of the electrons was diagnosed using a two-screen imaging magnetic spectrometer (see Methods).

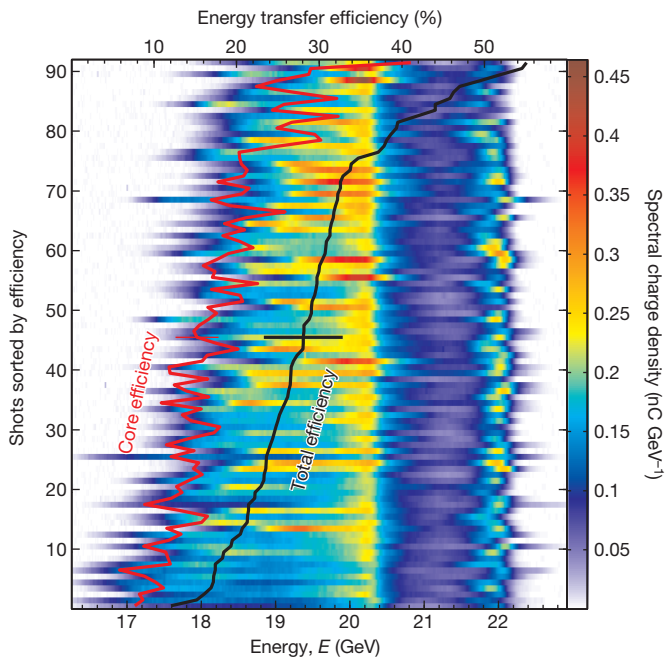
Figure 2 shows examples of electron spectra from three individual electron beam shots. The first shot, shown in Fig. 2a, gives the nominal spectrum of the two-bunch beam when the plasma source is not present. The second shot, shown in Fig. 2b, shows the measured spectrum when the plasma source is present and the spectrometer is set to optimally image 20.35-GeV electrons. This setting allows the energy loss of the drive bunch to be well characterized, while the trailing bunch, which has gained energy and has a higher divergence than the decelerated drive bunch, is strongly defocused. Of the approximately 1 nC of charge contained in the drive bunch entering the plasma, approximately  $450 \text{ pC} \pm 63 \text{ pC}$  ( $\pm \text{s.d.}$ ) appears at energies lower than 20.35 GeV. Figure 2c shows the spectrum of another shot similar to that shown in Fig. 2b but with the spectrometer set to image 22.35 GeV. Because the mean energy of the accelerated trailing bunch is close to the energy focus, it can readily be seen on the detector, whereas now the lowest-energy portion of the drive bunch has been strongly defocused, as expected. The centroid of the trailing bunch core in this shot has gained about 1.6 GeV of energy. The charge contained in the core of the accelerated portion of the spectrum is determined by an asymmetric Gaussian fit to the peak of the spectral projection for each shot (see Methods). The result of the core fit for the shot in Fig. 2c is shown in Fig. 2d as the dashed red line. Of the approximately 800 pC in the initial trailing bunch, the simulation indicates that only



**Figure 2 | Energetically dispersed beam profiles.** **a**, The dispersed electron beam profile without plasma interaction, where the spectrometer is set to image 22.35 GeV. Because the beam divergence is small, the entire spectrum of the beam is well resolved even at this imaging set point. **b** and **c**, The dispersed beam profile after the electron bunches have interacted with the plasma, where the spectrometer is set to image 20.35 GeV and 22.35 GeV, respectively:  $E_{\text{focus}}$  in **b** and **c**. The left  $x$  and top  $y$  axes correspond to the actual scale of the electron beam recorded at the spectrometer diagnostic screen, while the bottom  $E$  axis shows the calibrated energy axis along the dispersive dimension  $y$ . The scaling factors in **a–c** apply to the colour scale, quantifying transverse charge density. **d**, The spatially integrated spectrum (in  $x$ ) or the linear charge density of the bunches shown in **c** (solid blue line) along with the final spectrum obtained from the simulation depicted in Fig. 1b (solid green line in **d**). The core of the accelerated trailing beam is shown for the data (dashed red line).

about half of that resides in the first bucket of the wake. Of this, roughly  $200 \pm 33 \text{ pC}$  of charge appears above 20.35 GeV at the detection plane with about  $74 \pm 18 \text{ pC}$  of this contained in the core. The spread in charge quoted above refers to a data set containing 92 shots (discussed below). The projected energy spectrum (integrated over  $x$ ) for the shot shown in Fig. 2c is presented in Fig. 2d as the blue line. For comparison, the final spectrum of the simulated two-bunch plasma wakefield acceleration interaction shown in Fig. 1b—including transport losses and the focusing effects of the imaging spectrometer set to image at 22.35 GeV—is also plotted as the green line in Fig. 2d, and shows a good qualitative agreement with the experimental spectrum.

The full data set used in this analysis consists of 92 electron beam shots taken with the imaging spectrometer set to 22.35 GeV. The projected spectral profile for each shot was obtained and the resulting 92 spectra are plotted in Fig. 3 as a waterfall plot. The profiles are sorted according to the ‘total efficiency’ of energy transfer from the drive bunch to the trailing bunch via the wake. Here, the total efficiency (shown as the black line in Fig. 3) is defined as the net energy gain of all accelerated charge in the trailing bunch divided by the net energy loss of the drive bunch (see Methods). The mean total efficiency observed in the data set is 29.1% with a standard deviation of 8.9% and a maximum value of around 50%. For comparison, the total efficiency obtained from the simulation shown in Figs 1b and 2d is 49%. The core efficiency is shown in Fig. 3 as the red



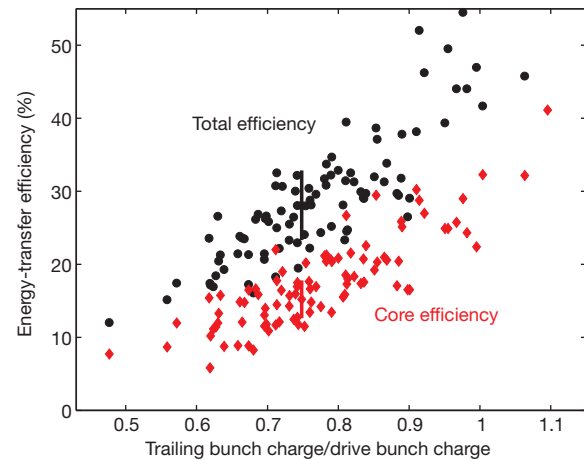
**Figure 3 | Spatially integrated electron beam spectra from the data set.** All 92 shots from the data set are shown with the imaging spectrometer set to image 22.35 GeV. The colour scale represents the energy spectrum in  $\text{nC GeV}^{-1}$ , similar to the blue curve of Fig. 2d, but transformed using the known dispersion of the spectrometer. The shots are sorted by the total energy-transfer efficiency (black line), as calculated using all charge above 20.35 GeV. The core energy-transfer efficiency, as calculated using only the charge found in the accelerated core of the trailing bunch, is shown as the red line. The black (or red) horizontal bar represents the typical systematic uncertainty of  $\pm 5\%$  (or  $\pm 3\%$ ) for the total energy-transfer efficiency (or core energy-transfer efficiency).

curve, which has a mean value of 17.7% with a standard deviation of 6.3%. It is lower than the total efficiency, as expected, owing to the smaller amount of charge contained in the core. The maximum core efficiency observed is over 30%.

We note that here we have measured a high energy-transfer efficiency from the drive bunch (via the wake) to the trailing bunch, though the drive bunch electrons lost only a small fraction of their total energy in our 36-cm-long plasma source. However, since the electrons in both bunches are ultra-relativistic, there is no relative motion between the bunches and the wake structure (see Fig. 1b and Supplementary Video 1), which means that the energy-transfer efficiency remains constant over the entire flat-density region of the plasma source (see Methods). Therefore, in a sufficiently long plasma source, nearly all of the energy in the drive bunch would be transferred to the trailing bunch with the same efficiency that we have measured. For the present experimental conditions, the length of the plasma would need to be extended to approximately 4 m in order to fully deplete the available energy in the drive bunch.

In this work, the average energy gain of the core of the trailing bunch is estimated to be  $1.6 \pm 0.1$  GeV. For our plasma length of about 36-cm full-width at half-maximum (FWHM), this represents an accelerating field gradient of  $4.4 \text{ GV m}^{-1}$ , which is in good agreement with the  $5 \text{ GV m}^{-1}$  seen in the simulation (red curve in Fig. 1b). The smallest energy spread of the trailing bunch core was 0.7% (compared to the initial energy spread of the trailing bunch of about 1%; see Methods), and the average energy spread was 2.0%, where the energy spread is defined as the r.m.s.  $\delta E/E$  for each shot. The simultaneous achievement of high gradient, high efficiency, and narrow energy spread demonstrated here represents an important advance in the development of a collider based on plasma wakefield acceleration<sup>19</sup>.

Increasing the peak current of the drive bunch by increasing its charge will in turn increase the amount of energy transferred from the drive



**Figure 4 | Energy-transfer efficiency dependence on wake loading.** The total efficiency (black circles) and core efficiency (red diamonds) versus the initial ratio of the charge in the trailing bunch to that in the drive bunch before entering the plasma source. The circles (diamonds) represent the same data that make up the black curve (red curve) in Fig. 3. The black (red) vertical bar represents the typical systematic uncertainty of  $\pm 5\%$  ( $\pm 3\%$ ) for the total energy-transfer efficiency (core energy-transfer efficiency). In this data, the total charge is held constant, and only the ratio of the charge of the trailing bunch to the charge of the drive bunch is varied.

bunch to the wake, thereby increasing the strength of the wakefield. This also increases the amount of charge needed in the trailing bunch to optimally load the wake and maximize the energy-transfer efficiency. We have used the natural spectral jitter of the incoming electron beam to measure this correlation between wake loading and energy transfer efficiency in our data.

Figure 4 shows a scatter plot of the shot-by-shot total and core efficiencies (black circles and red diamonds, respectively) obtained for the same data as in Fig. 3 plotted against the trailing-to-drive bunch charge ratio. A systematic error bar is shown for the same data point in each figure for both the total efficiency and the core efficiency values, where the leading source of uncertainty comes from the energy loss estimation (see Methods). The ratio of the charge of the trailing bunch to the charge of the drive bunch is used as a means of quantifying the loading of the wake by the trailing bunch normalized to the strength of the wake itself, which we here assume scales with the charge of the drive bunch. The observed correlation demonstrates the dependence of the energy-transfer efficiency on the loading of the wake and shows that a sufficiently high charge in the trailing bunch is necessary to attain a high efficiency. Although the final energy spread of the accelerated core of the trailing bunch is small at 2%, the increase from around 1% indicates that the shape of the trailing bunch profile inside the wake is not fully optimized for loading and flattening the wakefield, as shown in the simulation of Fig. 1b. The trailing bunch would need to be shorter in extent to more optimally flatten the wakefield.

In conclusion, high-efficiency acceleration of a distinct trailing bunch of electrons containing a substantial charge and having a small-energy spread has been demonstrated in a high-gradient, beam-driven plasma wakefield accelerator. These results bring plasma wakefield acceleration one step closer to becoming a viable accelerator technology.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 20 July; accepted 1 September 2014.**

- Chen, P., Dawson, J., Huff, R. & Katsouleas, T. Acceleration of electrons by the interaction of a bunched electron beam with a plasma. *Phys. Rev. Lett.* **54**, 693–696 (1985).
- Ruth, R., Chao, A., Morton, P. & Wilson, P. A plasma wake field accelerator. *Particle Accelerators* **17**, 171–189 (1985).

3. Esarey, E., Schroeder, C. B. & Leemans, W. P. Physics of laser-driven plasma-based electron accelerators. *Rev. Mod. Phys.* **81**, 1229–1285 (2009).
4. Joshi, C. & Katsouleas, T. Plasma accelerators at the energy frontier and on tabletops. *Phys. Today* **56**, 47–53 (2003).
5. Bingham, R., Mendonca, J. & Shukla, P. Plasma based charged-particle accelerators. *Plasma Phys. Contr. Fusion* **46**, R1 (2004).
6. Caldwell, A., Lotov, K., Pukhov, A. & Simon, F. Proton-driven plasma-wakefield acceleration. *Nature Phys.* **5**, 363–367 (2009).
7. Blumenfeld, I. *et al.* Energy doubling of 42 GeV electrons in a metre-scale plasma wakefield accelerator. *Nature* **445**, 741–744 (2007).
8. Lu, W., Huang, C., Zhou, M., Mori, W. B. & Katsouleas, T. Nonlinear theory for relativistic plasma wakefields in the blowout regime. *Phys. Rev. Lett.* **96**, 165002 (2006).
9. Huang, C. *et al.* QuickPIC: a highly efficient fully parallelized PIC code for plasma-based acceleration. *J. Phys. Conf. Ser.* **46**, 190–199 (2006).
10. An, W., Decyk, V. K., Mori, W. B. & Antonsen, T. M., Jr. An improved iteration loop for the three dimensional quasi-static particle-in-cell algorithm: QuickPIC. *J. Comput. Phys.* **250**, 165–177 (2013).
11. Hogan, M. *et al.* Multi-GeV energy gain in a plasma-wakefield accelerator. *Phys. Rev. Lett.* **95**, 054802 (2005).
12. Muggli, P. *et al.* Meter-scale plasma-wakefield accelerator driven by a matched electron beam. *Phys. Rev. Lett.* **93**, 014802 (2004).
13. Barov, N. *et al.* Ultra high-gradient energy loss by a pulsed electron beam in a plasma. *IEEE Proc. (2001 Particle Accelerator Conf.)* **1**, 126–128 (2001).
14. Katsouleas, T. Physical mechanisms in the plasma wake-field accelerator. *Phys. Rev. A* **33**, 2056–2064 (1986).
15. Tzoufras, M. *et al.* Beam loading in the nonlinear regime of plasma-based acceleration. *Phys. Rev. Lett.* **101**, 145002 (2008).
16. Rosenzweig, J. B. *et al.* Experimental observation of plasma wake-field acceleration. *Phys. Rev. Lett.* **61**, 98–101 (1988).
17. Muggli, P. *et al.* Photo-ionized lithium source for plasma accelerator applications. *IEEE Trans. (Plasma Sci.)* **27**, 791–799 (1999).
18. Green, S. Z. *et al.* Laser ionized preformed plasma at FACET. *Plasma Phys. Contr. Fusion* **56**, 084011 (2014).
19. Adli, E. *et al.* A beam driven plasma-wakefield linear collider: from Higgs factory to multi-TeV. In *Electronic Proceedings of the Snowmass 2013 Community Study on the Future of High-Energy Physics* <http://arxiv.org/abs/1308.1145> (2013).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** The FACET E200 plasma wakefield acceleration experiment was built and has been operated with funding from the United States Department of Energy. Work at SLAC was supported by DOE contract DE-AC02-76SF00515 and also through the Research Council of Norway. Work at UCLA was supported by DOE contracts DE-FG02-92-ER40727 and DE-SC0010064. Simulations were performed on the UCLA Hoffman2 and Dawson2 computers and on Blue Waters through NSF OCI-1036224. Simulation work at UCLA was supported by DOE contracts DE-SC0008491 and DE-SC0008316, and NSF contracts ACI-1339893 and PHY-0960344. The work of W.L. was partially supported by NSFC 11175102, the Thousand Young Talents Program and the Tsinghua University Initiative Scientific Research Program.

**Author Contributions** All authors contributed extensively to the work presented in this paper.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.L. (litos@slac.stanford.edu).



## METHODS

**Generation of two-bunch beam structure.** The SLAC linear accelerator provides a 20.35-GeV electron beam with a charge of 3.2 nC to the FACET experimental area. To create a two-bunch structure with a peak-to-peak separation of  $\sim 135 \mu\text{m}$ , a single high-charge bunch with a roughly linear head-to-tail correlated energy spread is energetically dispersed and physically segmented in a plane transverse to its direction of motion.

Extended Data Fig. 1 shows a schematic of the FACET experimental area. The five dipole bend magnets at the vertices of the 'W'-shaped beam line form a chicane, which allows for the manipulation of correlations between an electron's momentum  $p_z$  and its longitudinal position  $z$  within the bunch. In inset 1, the red-to-blue colour transition indicates the initial longitudinal energy correlation generated in the linear accelerator upstream of the experimental area. Higher energies (red) are at the head of the bunch, and there is no transverse ( $p_z, x$ ) correlation at this point. After some propagation beyond the first dipole magnet in the chicane, the beam becomes highly dispersed in the transverse plane with a strong linear ( $p_z, x$ ) correlation (inset 2). At this point, a 305- $\mu\text{m}$ -wide, 16-mm-thick tantalum bar (the 'notching device' in Extended Data Fig. 1a), is inserted into the middle of the  $\sim 2$ -mm-wide beam. Electrons that intercept this tantalum bar lose energy through collisions and are scattered at large angles, causing them to be ejected from the beam line by the strong magnets further downstream in the chicane. This process removes  $\sim 40\%$  of the initial charge in the beam (inset 3). At a symmetric location before the final dipole of the chicane (inset 4), the ( $p_z, x$ ) correlation of the bunch returns with the opposite sign. Here, a compact magnetic wiggler (Extended Data Fig. 1c) yields synchrotron X-rays that sweep across a scintillating YAG screen, creating a profile of the notched electron spectrum. This is the 'initial spectrometer', providing an optical replica of the notched electron beam spectrum. Inset 5 shows that after leaving the chicane the bunch is 'over-compressed', that is, the lower energy particles are in the front of the now segmented beam, comprising the drive bunch, while the higher energy particles are in the rear, comprising the trailing bunch. The longitudinal profile of the two-bunch structure is measured in the time domain with a transverse deflecting X-band radio-frequency structure (Extended Data Fig. 1b) that streaks the beam onto a profile monitor screen located near the plasma source.

**Electron beam characteristics.** The electron beam sent into the plasma source has an incoming energy of 20.35 GeV with a FWHM spread of  $\pm 2\%$ , and a total charge of 1.80 nC. The drive bunch contains 1.02 nC, with a r.m.s. length of  $\sigma_z = 25 \mu\text{m}$ , and peak current  $I_{\text{peak}} = 4.9 \text{ kA}$ . The broader trailing bunch contains 780 pC, with  $\sigma_z = 47 \mu\text{m}$ , and  $I_{\text{peak}} = 2.0 \text{ kA}$ . The trailing bunch also exhibits a correlated energy spread of about 1% r.m.s. The peak-to-peak separation of the two bunches is 135  $\mu\text{m}$ . After exiting the chicane and before entering the plasma source, the beam is focused by a series of five quadrupole magnets, labelled the 'final focus quadrupole magnets' (Extended Data Fig. 1d), down to an r.m.s. transverse spot size of 30  $\mu\text{m}$  in the middle of the plasma density up-ramp. The normalized emittance of the beam in the transverse dimensions is roughly  $\epsilon_{n,x} = 358 \text{ mm mrad}$  and  $\epsilon_{n,y} = 35.8 \text{ mm mrad}$ . Extended Data Fig. 2 shows the longitudinal profile of a typical two-bunch beam used in the experiment, as measured by the transverse deflecting radio-frequency structure.

**Plasma source.** The plasma source used in this experiment is a laser-ionized column of lithium vapour contained inside a heat pipe oven<sup>17,18</sup> with a uniform density of  $5.0 \times 10^{16} \text{ cm}^{-3}$  over a 26-cm-long region with 10-cm density ramps on either side, giving a FWHM length of approximately 36 cm. The lithium vapour column is depicted in Extended Data Fig. 1f. Extended Data Fig. 3 shows the density profile of the neutral vapour pressure density deduced from the measured temperature profile along the oven containing the lithium as well as the simple fit used to describe the density profile in our model. The density and length of the lithium vapour is controlled through the temperature of the oven and the pressure of a room-temperature noble gas (argon) serving as a buffer for the lithium at either end of the pipe. A 200-fs-long Ti:sapphire laser pulse (Extended Data Fig. 1e) containing 250 mJ of energy is focused by a 1.5° axicon lens that produces a zero-order Bessel beam profile through the full length of the lithium vapour, which in turn creates a plasma column  $\sim 1 \text{ mm}$  in diameter<sup>18</sup>. The laser arrives 100 ps before the arrival of the electron beam, which is over an order of magnitude earlier than substantial recombination of the plasma is expected to occur. Calculations indicate that the peak intensity of the laser is sufficient to ionize a  $\sim 20$ - $\mu\text{m}$ -diameter filament in the buffer gas for up to 30 cm upstream of the lithium vapour. Interaction of the beam with this filament may lead to some nonlinear focusing of the beam, preventing some of the electrons in the incoming beam from cleanly coupling into the lithium plasma and thus preventing them from participating in the experiment.

**Electron imaging spectrometer.** The spectrum of the electrons exiting the plasma is diagnosed with an imaging spectrometer having two diagnostic screens. An imaging quadrupole doublet, depicted in Extended Data Fig. 1g, is required to capture and deliver the electrons over the long distance between the exit of the plasma and the two screens. A strong dipole magnet, in Extended Data Fig. 1h, vertically disperses

the electrons onto the two screens, shown as Extended Data Fig. 1i. A camera viewing one screen records Cherenkov light produced by the beam in a 1.4-cm air gap between two silicon wafers, while another camera records the scintillation light produced by a phosphor screen after the electron beam has passed through it. The imaging condition at the Cherenkov screen for the analysed data shown in Figs 2c, 3 and 4 is for an energy of 22.35 GeV in both the horizontal and vertical (dispersive) plane. The imaging condition at the phosphor screen, located 1 m upstream of the Cherenkov screen, is for 22.25 GeV in the horizontal plane, and 21.50 GeV in the vertical plane. The combination of these different imaging conditions provided confidence that the signal observed in the data was not simply the result of an enhancement introduced by the focusing condition of the beam.

The spatial resolution of the spectrometer in the energy plane is dominated by scattering of the beam in the vacuum-to-air exit window. The r.m.s. multiple scattering angle due to collisions of the beam particles within the solid exit window is calculated to be 143  $\mu\text{rad}$ , using standard formulae. The profile monitor is located 95 cm downstream of the exit window, yielding a contribution to the spatial resolution of 135  $\mu\text{m}$ . The measured spatial resolution is about 150  $\mu\text{m}$ . We define the energy resolution as the spatial resolution divided by the dispersion, times the beam energy. The dispersion induced by the spectrometer dipole, at the nominal FACET beam energy of 20.35 GeV, is 62 mm. This gives an energy resolution of 76 MeV at the nominal energy of 20.35 GeV. The energy resolution scales as energy squared, yielding an energy resolution of 91 MeV for the highest imaging energy referred to in the paper, 22.35 GeV. The total distance from the object plane to the imaging plane is 22.6 m, and the imaging properties of the spectrometer shows therefore little sensitivity to the position of the object plane on the order of 10 cm or less. For example, calculations show that the error in the imaged energy is on the order of 1% or less for particles with an angle of up to 1 mrad at the object plane and for errors of the position of the object plane up to  $\pm 10 \text{ cm}$ .

The beam size after the spectrometer is imaged with a magnification of 0.5 in the energetically dispersed ( $y$ ) plane, and the spot size in this dimension is thus dominated by the energy spread. In the non-dispersed ( $x$ ) plane, the spectrometer magnifies by a factor of 5.3, so the beam size as it appears on our diagnostic screen and as it is plotted in Fig. 2 is larger than it is at the plasma exit plane by this factor.

**Plasma wakefield acceleration simulation.** The simulations presented in Fig. 1 and Fig. 2d use the quasi-static, three-dimensional, particle-in-cell code QuickPIC<sup>9,10</sup>. The input beam is Gaussian in both transverse and longitudinal dimensions, with r.m.s. sizes and emittance values corresponding to those measured for the experimental beam, as listed in the 'Electron beam characteristics' section. The simulated beam has zero initial energy spread and the initial charge in the simulation corresponds to roughly 75% of the charge measured upstream of the plasma source in the experiment, as this yields the best final spectrum agreement with the data in Fig. 2d. The reduced charge in the simulation is probably compensating for electrons in the experimental beam that are unable to couple into the plasma wake owing to potential factors not captured in the simulation, such as non-Gaussian tails in the charge distribution. The simulated plasma density profile was based on the measured plasma source profile: a 26-cm-long flat-top density of  $5.0 \times 10^{16} \text{ cm}^{-3}$  with 10-cm-long density ramps on each side, giving a FWHM length of 36 cm.

For this simulation run, a snapshot of the particles and fields was recorded 120 times over a 51-cm propagation distance (including transition into and out of the plasma density ramps on either end), thus fully resolving the oscillations of the envelope of the electron bunch which occurs on approximately a centimetre scale. The simulation box tracks the beam-plasma interaction in the coordinates  $x, y, \zeta = z - ct$ ; that is, the box moves at light speed, close to the velocity of the bunches, although it sees the beam and plasma in the reference frame of the laboratory. The box has a size of  $601 \mu\text{m} \times 601 \mu\text{m} \times 481 \mu\text{m}$  in the two transverse dimensions and the longitudinal dimension, respectively. The number of the cells for the simulation box is  $512 \times 512 \times 512$  ( $\sim 134$  million cells in total).

The full, 120-frame simulation movie from which Fig. 1b was taken is available as Supplementary Movie 1. It depicts the evolution of the two-bunch beam structure as it propagates through the plasma. The upper portion of the frame shows the charge density of the beam and the plasma wake, and the lower portion of the frame shows the evolution of the beam's energy as a function of the longitudinal position inside the plasma wake.

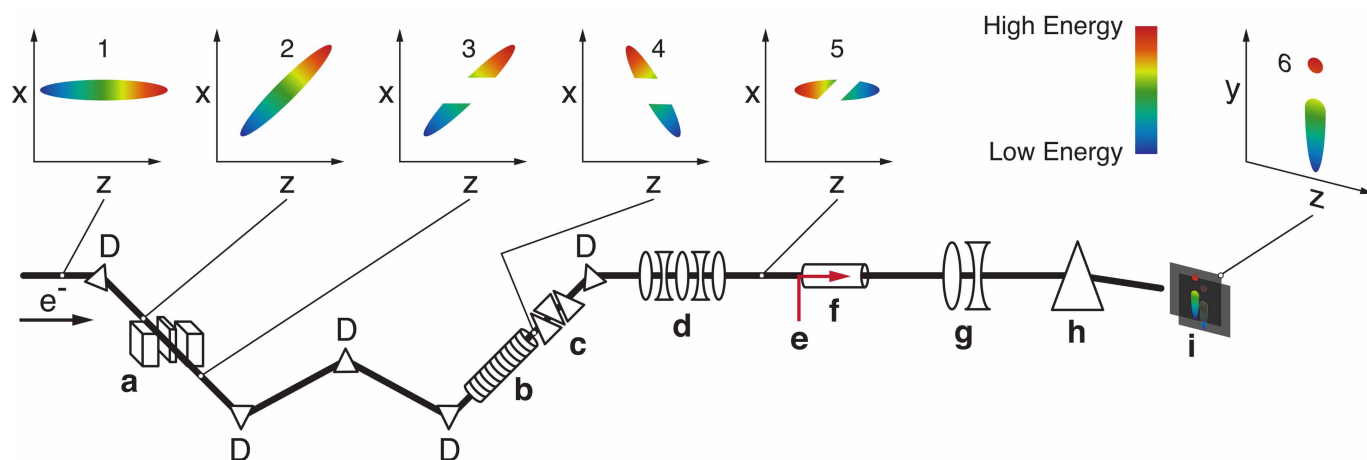
In the simulation, the drive bunch can be seen to be scalloped because the ion channel is not formed instantaneously within the beam. A slice at the front of the beam feels no focusing force while later slices feel a progressively stronger force (progressively higher betatron oscillation frequency) up to the slice that resides in the fully formed ion channel. A new scallop forms after each full betatron oscillation for the slices in the ion channel<sup>7</sup>. The curvature of the scalloping is due to a nonlinear dependence of the focusing force during the rise time of the bunch when the ion column is still forming. A similar scalloping structure is also observed in the experiment, as seen on the energy loss portion of the drive bunch spectrum shown in Fig. 2b.

**Transfer efficiency, energy spread, and core charge estimation.** The net energy gain for a particular shot is calculated as  $E_{\text{gain}} = \sum_{E_i > E_0} (E_i - E_0) q_i$ , where  $E_i$  is the calibrated beam energy corresponding to pixel row  $i$  on the charge-coupled device (CCD) image,  $E_0 = 20.35$  GeV is the initial beam energy, and  $q_i$  is the total amount of charge observed in pixel row  $i$ . The energy loss may be calculated in the same way, though the energy loss is known to be under-represented in the 92 shots analysed in Figs 3 and 4 owing to the energy setting of the imaging quadrupoles. Thus, we instead use an estimation of the energy loss for each shot that is derived from an empirical quadratic correlation (normalized residual is 0.87) between the energy loss observed on the Cherenkov detector and the initial drive bunch charge measured upstream of the plasma source for a data set taken when the quadrupole imaging condition was set to 20.35 GeV (as in Fig. 2b) immediately before the primary 92-shot data set. For these conditions, the decelerated drive bunch could be well observed on the Cherenkov detector and the net energy loss well quantified.

To determine the energy gain of the core of the accelerated bunch, a fit was performed to the projected spectrum of each shot in the data set. Figure 2c shows a dispersed beam image of a single shot with the imaging spectrometer set to image 22.35 GeV. The spectral projection of this shot, shown in Fig. 2d, indicates a narrow peak near 22 GeV, and a more diffuse, continuous distribution of accelerated charge with a higher divergence that extends down to the initial beam energy. A fit to the projected spectrum above 21 GeV is performed that is the sum of two distinct parts: an asymmetric Gaussian to characterize the peak that occurs near 22 GeV, and a

half-Gaussian to account for the diffuse, high-divergence charge. The asymmetric Gaussian portion of the fit is then used to quantify the properties of the accelerated core of the trailing bunch, such as the charge, energy gain, energy spread and energy-transfer efficiency. Extended Data Fig. 4 shows the fit that was used for the shot displayed in Fig. 2c and d, where the red line indicates the core of the accelerated trailing bunch. The same function describing the core is also shown as the dashed red line in Fig. 2d.

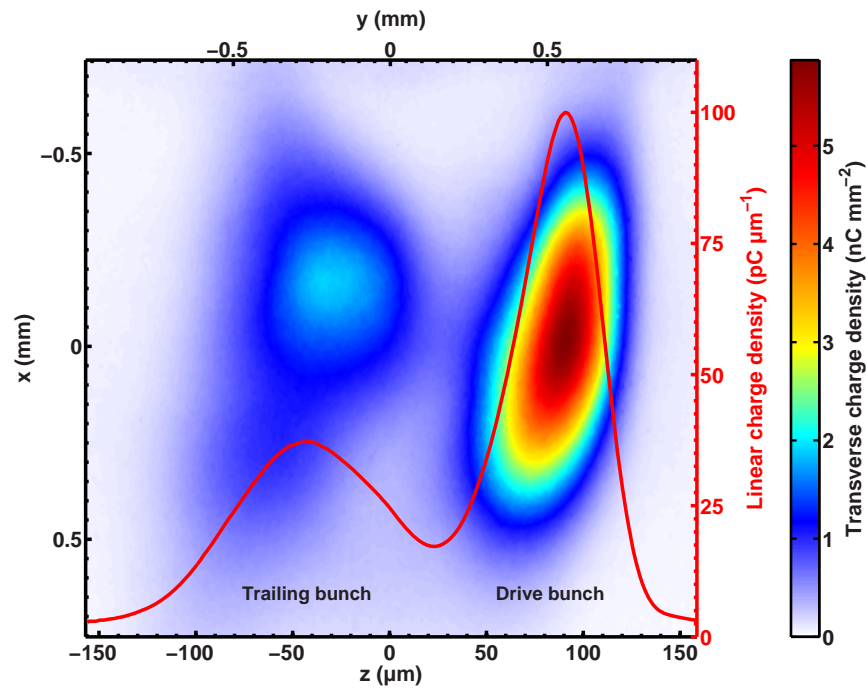
The 92 shots analysed in the main paper come from two data sets of 50 shots each, taken within minutes of one another, while the imaging spectrometer was set to image an energy of 22.35 GeV. Of the total 100 combined shots, 8 were rejected as outliers in which the beam-plasma interaction was substantially weaker than in the remaining 92 shots. This was quantified by the amount of ‘non-participating charge’, that is, the amount of charge found within a  $\pm 2\%$  energy window about the beam’s initial energy of 20.35 GeV (corresponding to the initial beam’s FWHM energy spread), and by the energy value above the initial beam energy with the greatest charge density, or the ‘peak energy’ of the accelerated electrons. A two-dimensional cut on peak energy and non-participating charge is applied to the data that rejects the 8 low interaction shots. Five of these shots had the lowest drive-bunch charge (and thus the lowest peak current) of the complete data set. A low-peak-current drive bunch impedes the ability to form a wake in the blowout regime, leading to a weak interaction in the plasma. All eight of the rejected shots may also have had large transverse sizes, which can similarly impede the ability to form a strong wake.



**Extended Data Figure 1 | FACET experimental area schematic.** Electron beam line features: **a**, beam notching device, **b**, transverse deflecting structure, **c**, initial spectrometer, **d**, final-focus quadrupole magnets, **e**, lithium plasma ionization laser, **f**, lithium vapour column, **g**, spectrometer imaging quadrupole

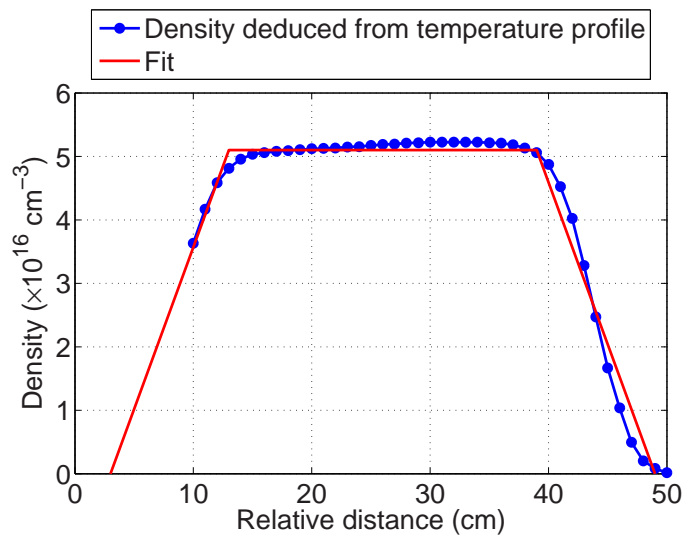
magnets, **h**, spectrometer dipole magnet, and **i**, Cherenkov and phosphor screens. Bend dipole magnets in the 'W'-shaped chicane are each labelled 'D'. The arrow beneath the  $e^-$  symbol indicates the electron beam's direction of motion (left to right).



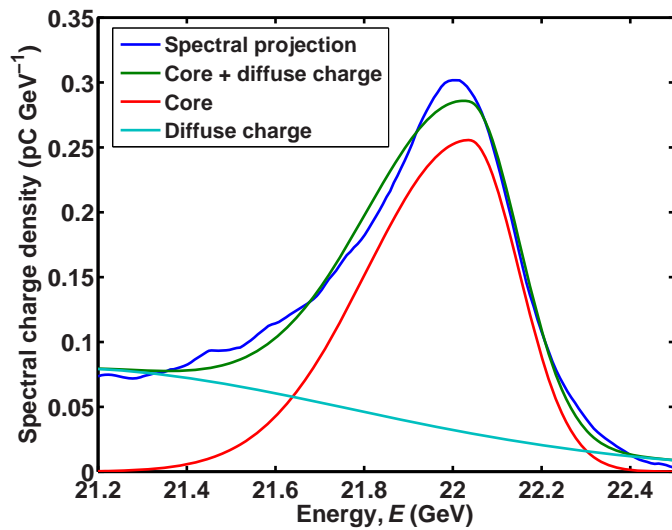


**Extended Data Figure 2 | Measured longitudinal profile of two-bunch beam.** Image of a typical two-bunch beam streaked onto a profile monitor screen by the transverse deflecting radio-frequency structure (Extended Data Fig. 1b). The drive bunch appears on the right-hand side. Overlaid on the image is the projected longitudinal profile (red line). The left ( $x$ ) and top ( $y$ ) axes show the transverse dimensions of the streaked beam on the profile

monitor screen, while the colour axis indicates the charge density of the transverse profile. The bottom ( $z$ ) axis shows the streaked dimension ( $y$ ) with the appropriate scaling factor applied to give the corresponding longitudinal coordinate. The right axis shows the linear charge density corresponding to the projected longitudinal profile.



**Extended Data Figure 3 | Lithium vapour column density profile.** The profile of the neutral vapour pressure density of the lithium vapour column deduced from the measured temperature profile (temperature versus relative distance of insertion of a thermocouple probe) along the heat pipe oven is shown as the blue line. The simple fit used to describe the density profile in our model is shown as the red line.



**Extended Data Figure 4 | Fit to accelerated charge.** The blue line is the spectral projection of the same data shot shown in Fig. 2c and d. The green line is a fit to the data using a half-Gaussian tail (cyan line) to account for the diffuse, high-angular-divergence accelerated charge plus a full, asymmetric Gaussian (red) used to describe the core of the accelerated trailing bunch after subtracting the half-Gaussian tail.



# Solution-processed, high-performance light-emitting diodes based on quantum dots

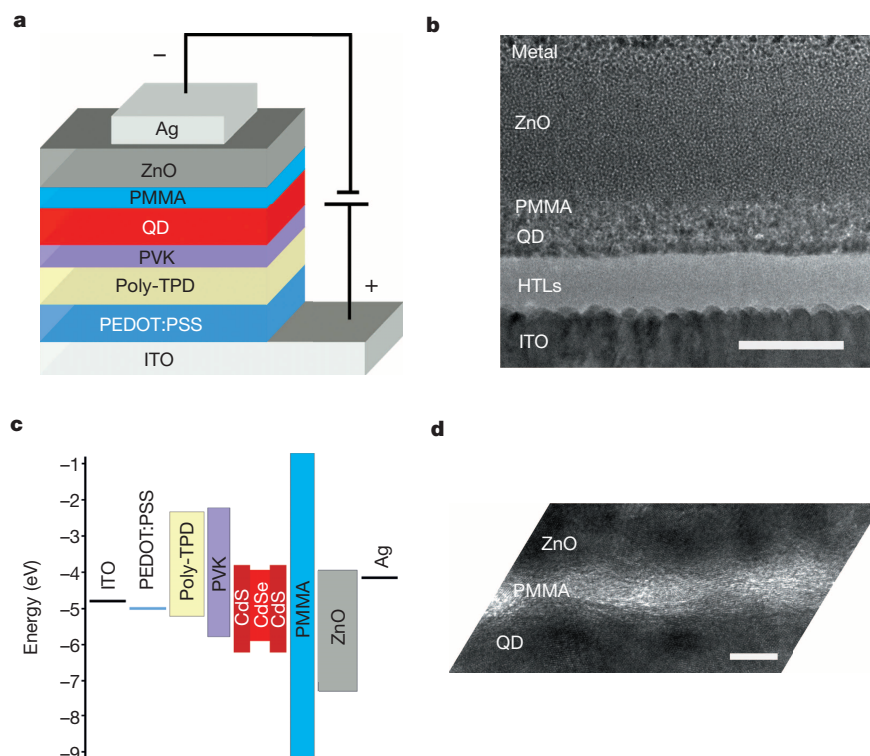
Xingliang Dai<sup>1</sup>, Zhenxing Zhang<sup>2</sup>, Yizheng Jin<sup>1</sup>, Yuan Niu<sup>2</sup>, Hujia Cao<sup>2</sup>, Xiaoyong Liang<sup>1</sup>, Liwei Chen<sup>3</sup>, Jianpu Wang<sup>4</sup> & Xiaogang Peng<sup>2</sup>

Solution-processed optoelectronic and electronic devices are attractive owing to the potential for low-cost fabrication of large-area devices and the compatibility with lightweight, flexible plastic substrates. Solution-processed light-emitting diodes (LEDs) using conjugated polymers or quantum dots as emitters have attracted great interest over the past two decades<sup>1,2</sup>. However, the overall performance of solution-processed LEDs<sup>2–5</sup>—including their efficiency, efficiency roll-off at high current densities, turn-on voltage and lifetime under operational conditions—remains inferior to that of the best vacuum-deposited organic LEDs<sup>6–8</sup>. Here we report a solution-processed, multilayer quantum-dot-based LED with excellent performance and reproducibility. It exhibits colour-saturated deep-red emission, sub-bandgap turn-on at 1.7 volts, high external quantum efficiencies of up to 20.5 per cent, low efficiency roll-off (up to 15.1 per cent of the external quantum efficiency at 100 mA cm<sup>−2</sup>), and a long operational lifetime of more than 100,000 hours at 100 cd m<sup>−2</sup>, making this device the best-performing solution-processed red LED so far, comparable to state-of-the-art vacuum-deposited organic LEDs<sup>2–8</sup>. This optoelectronic performance is achieved by inserting an insulating layer between the quantum dot layer and the oxide electron-transport layer

to optimize charge balance in the device and preserve the superior emissive properties of the quantum dots. We anticipate that our results will be a starting point for further research, leading to high-performance, all-solution-processed quantum-dot-based LEDs ideal for next-generation display and solid-state lighting technologies.

Quantum dots are solution-processable semiconductor nanocrystals<sup>9–11</sup> that promise size-tunable emission wavelengths, narrow emission linewidths, near-unity-photoluminance quantum yield and inherent photophysical stability. As inorganic crystalline emission centres, quantum dots are expected to be promising candidates to overcome stability problems of both polymer LEDs and small-molecule organic LEDs (OLEDs), such as drastic efficiency roll-off at high current densities and low operational lifetime. To fully exploit the superior properties of quantum dots, a number of quantum-dot-based LED (QLED) structures were developed and various materials, including small molecules, conjugated polymers and inorganic oxides, were explored as charge-transport interlayers<sup>3,12–20</sup>.

Our device (Fig. 1a, b) consists of multiple layers of, in the following order, indium tin oxide (ITO), poly(ethylenedioxythiophene):polystyrene sulphonate (PEDOT:PSS, 35 nm), poly(*N,N'*-bis(4-butylphenyl)-*N,N'*-bis(phenyl)-benzidine) (poly-TPD, 30 nm), poly(9-vinylcarbazole)



**Figure 1 | Multilayer QLED device.** **a**, Device structure. **b**, Cross-sectional transmission electron microscopy image showing the multiple layers of material with distinct contrast. Scale bar, 100 nm. The PMMA layer is evident only when the cross-sectional sample is sufficiently thin (**d**) because the neighbouring quantum dot layer and the ZnO layer can obstruct the imaging of the PMMA layer. HTL, hole-transport interlayer. **c**, Flat-band energy level diagram. **d**, High-magnification transmission electron microscopy image of an extremely thin cross-sectional sample revealing the presence of the PMMA layer between the ZnO layer and the quantum dot layer. Scale bar, 5 nm.

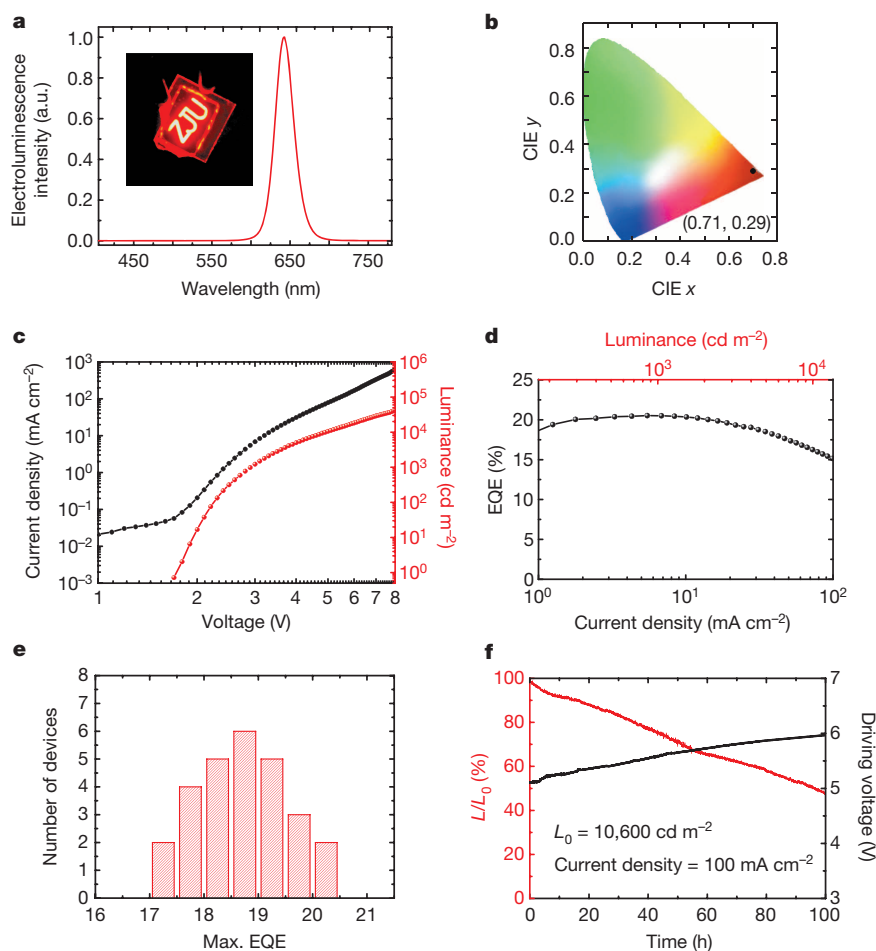
<sup>1</sup>Center for Chemistry of High-Performance & Novel Materials, State Key Laboratory of Silicon Materials, Cyrus Tang Center for Sensor Materials and Applications, School of Materials Science and Engineering, Zhejiang University, Hangzhou 310027, China. <sup>2</sup>Center for Chemistry of High-Performance & Novel Materials, Department of Chemistry, Zhejiang University, Hangzhou 310027, China. <sup>3</sup>i-Lab, Suzhou Institute of Nano-Tech and Nano-Bionics, Chinese Academy of Sciences, 398 Ruoshui Road, Suzhou Industrial Park, Suzhou 215123, China. <sup>4</sup>Key Laboratory of Flexible Electronics (KLOFE) & Institute of Advanced Materials (IAM), National Jiangsu Synergistic Innovation Center for Advanced Materials (SICAM), Nanjing Tech University (NanjingTech), 30 South Puzhu Road, Nanjing 211816, China.

(PVK, 5 nm), CdSe–CdS core–shell quantum dots (QDs, 40 nm), poly (methyl methacrylate) (PMMA, 6 nm), ZnO nanoparticles (150 nm) and silver (Ag, 100 nm). Figure 1c shows a schematic of the flat-band energy level diagram of the layers. The energy level values for ITO, PEDOT:PSS and ZnO were obtained by ultraviolet photoelectron spectroscopy and optical measurements. Other energy level values were taken from refs 3, 21–23. Phase-pure zinc blende CdSe–CdS core–shell nanocrystals with ten monolayers of CdS shell are used. These quantum dots possess a photoluminance quantum yield of >90% and outstanding optical properties<sup>24</sup> (Extended Data Fig. 1). Bilayer-structured hole-transport interlayers of poly-TPD/PVK take advantage of the deep highest-occupied-molecular-orbit energy level of PVK to realize efficient hole injection into the quantum dot layers and the relatively high hole mobility of poly-TPD to achieve low turn-on voltage and high power efficiency. Thin films of colloidal ZnO nanocrystals (Extended Data Fig. 2) are employed as electron-transport interlayers (ETLs) because of their unique combination of high electron mobility, ease of preparation and the previously identified benefit of efficient electron injection into the quantum dot layers<sup>17,18</sup>. The key component of this device, a thin insulating PMMA layer (Fig. 1d), is inserted between the ZnO ETL and the quantum dot emissive layer.

The normalized electroluminescence spectrum of the QLED is shown in Fig. 2a. The symmetric emission peak at 640 nm with a narrow full-width at half-maximum of 28 nm corresponds to Commission Internationale de l'Eclairage (CIE) colour coordinates of (0.71, 0.29), which are close to the spectral locus and represent colour-saturated deep-red

emission ideal for display applications (Fig. 2b). Figure 2c shows the current density–voltage and luminance–voltage characteristics of a device with the best efficiency. The current density and luminance increase steeply once the voltage reaches  $\sim 1.7$  V, yielding a maximum brightness of over  $42,000 \text{ cd m}^{-2}$  at 8 V. The peak external quantum efficiency (EQE), 20.5%, is achieved at a current density of  $\sim 7 \text{ mA cm}^{-2}$  and a brightness of  $\sim 1,200 \text{ cd m}^{-2}$ . The peak EQE of this device, 20.5%, is the highest value for QLEDs<sup>3,12–20</sup>. High EQE can be maintained in a wide range of current densities (Fig. 2d), that is, EQE > 18% when the current density is in the range of 1–42  $\text{mA cm}^{-2}$ , which corresponds to a brightness in the range of 100–6,600  $\text{cd m}^{-2}$ . When the current density reaches  $100 \text{ mA cm}^{-2}$ , an EQE of >15% is sustained. The low efficiency roll-off of this device, which is better than that of the other high-efficiency QLED (peak EQE, 18.5%) with vacuum-deposited hole-transport interlayers<sup>3</sup> and is comparable to those of state-of-the-art vacuum-deposited OLEDs<sup>8</sup>, suggests that our QLEDs are promising for high-power applications. The established solution-processing protocol leads to devices with excellent reproducibility. As shown by the histograms for 27 devices from four batches (Fig. 2e), both the high average peak EQE, 18.7%, and the low relative standard deviation of peak EQE, 4.3%, are encouraging.

Our QLEDs, simply sealed by ultraviolet-curable resin and without other complicated encapsulation techniques, exhibit outstanding ambient stability under high-brightness conditions. As shown in Fig. 2f, for a typical device tested at a constant driving current density of  $100 \text{ mA cm}^{-2}$ , which corresponds to an initial luminance,  $L_0$ , of  $10,600 \text{ cd m}^{-2}$ , the half-lifetime,  $T_{50}$ , defined as the time for the luminance to decrease to  $L_0/2$ ,



**Figure 2 | Device performance.** **a**, Electroluminescence spectrum at an applied voltage of 3 V and, inset, a photograph of a device with the Zhejiang University logo. a.u., arbitrary units. **b**, The corresponding CIE coordinates. **c**, Current density and luminance versus driving voltage characteristics for the device with the best efficiency. **d**, EQE versus current density and luminance

for the device with the best efficiency. **e**, Histogram of peak EQEs measured from 27 devices. **f**, Stability data for a QLED device ( $L$ , luminance). The device was test at ambient conditions (temperature, 20–25 °C; relative humidity, 50–70%).

is 95 h. By using the relation  $L_0^n T_{50} = \text{const.}$  and assuming an acceleration factor of  $n = 1.5$  (ref. 25),  $T_{50}$  for this device at  $100 \text{ cd m}^{-2}$  is predicted to be over 100,000 h. The remarkable operational stability of our device, along with its outstanding efficiency, low efficiency roll-off, sub-bandgap turn-on voltage and excellent reproducibility, marks a milestone in the production of QLEDs for practical applications.

For solution-processed LEDs, layer-by-layer deposition of high-quality films without intermixing is essential for constructing high-performance devices. In our QLEDs, six layers, including PEDOT:PSS, poly-TPD, PVK, quantum dots, PMMA and ZnO, are deposited from solution. The top Ag electrodes are fabricated by vacuum deposition. Atomic force microscopy and scanning Kelvin probe microscopy analyses show that the six solution-processed layers have pin-hole-free features (Extended Data Fig. 3) and evenly distributed surface potentials (within  $\pm 40 \text{ meV}$ ; Extended Data Fig. 4). Confocal microscopy imaging reveals that the quantum dot layers have homogeneously emissive properties. Specifically, optical measurements (Extended Data Fig. 5) indicate an average thickness of  $\sim 6 \text{ nm}$  for the PMMA layer in the QLEDs. The root mean squared roughness of the quantum dot film is in the range of  $1.6\text{--}2.6 \text{ nm}$ . After depositing the PMMA layer, the root mean squared roughness decreases to  $0.6\text{--}1.6 \text{ nm}$ . On the basis of the above facts, we suggest that the PMMA layer on the quantum dot film is continuous, and that the local thickness of the PMMA layer fluctuates owing to the relatively rough quantum dot surface (Fig. 1d).

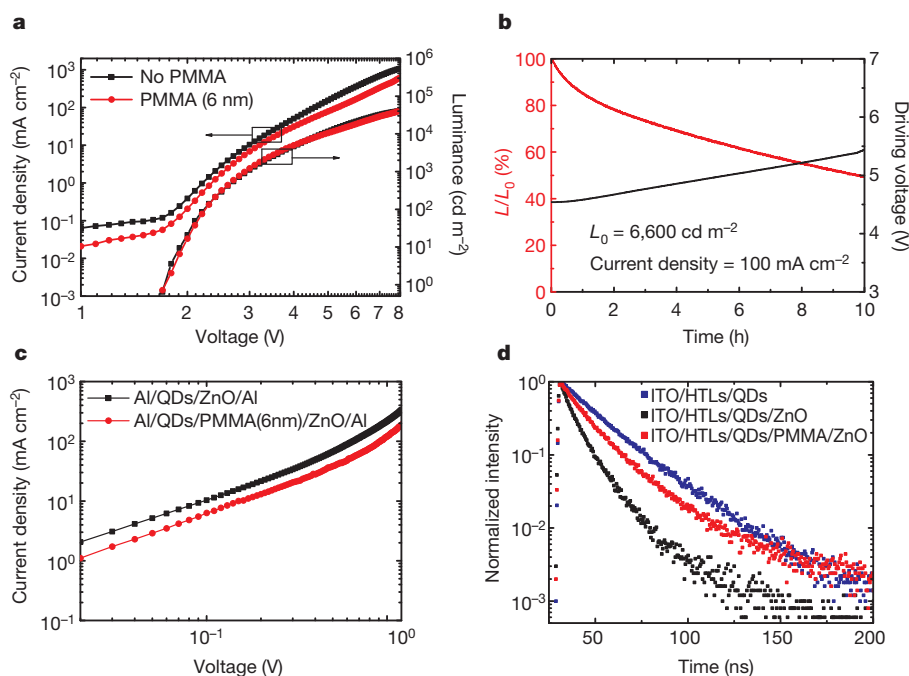
We highlight that the incorporation of the insulating PMMA layer with a suitable thickness between the ZnO ETL and the quantum dot emissive layer optimizes charge balance in the device. In our device, there is a moderate energetic barrier for hole injection owing to the deep valance-band energy level of the quantum dots (Fig. 1c). Furthermore, the hole mobility of poly-TPD<sup>21</sup> ( $1 \times 10^{-4} \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$ ) and PVK<sup>22</sup> ( $2.5 \times 10^{-6} \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$ ) are one to three orders of magnitude lower than the electron mobility of ZnO nanocrystal films ( $\sim 1.8 \times 10^{-3} \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$ ) (Extended Data Fig. 2b). These factors can lead to excess electron injection into the quantum dot emissive layer. The unbalanced charge injection in the QLEDs is reflected by the significantly higher current densities of the electron-only devices (ITO/Al/QDs/ZnO/Al) than those of the

hole-only devices (ITO/PEDOT:PSS/poly-TPD/PVK/QDs/Pd) (Extended Data Fig. 6).

The excess electron current in the QLED without the PMMA layer degrades device performance. As shown in Fig. 3a, for the QLED without the PMMA layer, both the turn-on voltage and the brightness in the entire range of forward bias are nearly the same as for the device with the  $6 \text{ nm}$  PMMA layer, whereas the current densities are much greater. Therefore, the excess electron current substantially lowers the efficiency of the device. Furthermore, the QLEDs without the PMMA layers exhibit relatively poor stability. As shown in Fig. 3b, when tested at a constant driving current density of  $100 \text{ mA cm}^{-2}$ ,  $T_{50}$  for this device with an  $L_0$  of  $6,600 \text{ cd m}^{-2}$  is only 10 h. Converting to the values at the same initial brightness of  $100 \text{ cd m}^{-2}$ , this is only  $\sim 5\%$  of that for the device with the  $6 \text{ nm}$  PMMA layer. Therefore without efficient radiative recombination to release the energy, the excess electron current can deteriorate the QLEDs rapidly under operational conditions.

It is possible to modulate the electron injection from the ZnO ETLs to the quantum dot layers and eliminate excess electron currents in the QLEDs by adjusting the thickness of the inserted PMMA layers. When a PMMA layer is applied in the electron-only devices (ITO/Al/QDs/PMMA/ZnO/Al; Fig. 3c and Extended Data Fig. 7a), the current density gradually reduces as the thickness of the PMMA layer increases.

For the working devices, a gradual decrease in current density on the increase of the thickness of the PMMA layers is also observed (Fig. 3a and Extended Data Fig. 7b). For the device with the  $6 \text{ nm}$  PMMA layer, a peak EQE of 20.5% (Fig. 2d), corresponding to a close-to-unity internal quantum efficiency (estimated to be 88.6% by assuming a Lambertian emission profile<sup>26</sup>), indicates almost perfect charge balance in this device. This is consistent with the low efficiency roll-off and improved operational stability of the QLEDs with the  $6 \text{ nm}$  PMMA layers (Fig. 2d, f). Further increasing the thickness of the PMMA layers to  $> 6 \text{ nm}$  results in both increase of the turn-on voltage and decrease of the brightness (Extended Data Fig. 7c, d). These control experiments, along with the results described above, clearly suggest that charge balance in the QLEDs can be optimized by inserting an insulating layer with a suitable thickness. Either excess electron injection or over-blocking electron current



**Figure 3 | Impacts of the 6 nm PMMA layer.** **a**, Current density and luminance versus voltage characteristics for QLEDs without and with the  $6 \text{ nm}$  PMMA layer. **b**, Stability data for a QLED without the PMMA layer. **c**, Current density–voltage curves for the electron-only devices showing that the  $6 \text{ nm}$  PMMA layer results in a  $\sim 1.8$ -fold decrease in current density. The thicknesses

of the ZnO layer and the quantum dot layer are 150 and  $40 \text{ nm}$ , respectively. **d**, Time-resolved photoluminescence decay for the quantum dot films contacting different layers. The thicknesses of the layers are identical to those in the optimized QLED.



deteriorates charge balance in the QLEDs and thereby degrades device performance.

For both photoluminescence<sup>27</sup> and electroluminescence<sup>19</sup>, charging degrades the emissive properties of the quantum dots. In our case, when the quantum dots are in direct contact with the ZnO ETLs, a spontaneous charge transfer process occurs owing to the work function difference, leaving positively charged quantum dots<sup>3</sup>. Charging of the quantum dots causes inefficient trion emissions<sup>27</sup>, as indicated by the fact that the average photoluminescence lifetime of the quantum dot films decreased from 21.6 to 10.6 ns after the deposition of the top ZnO films (Fig. 3d). The insertion of a thin PMMA layer modifies the QD/ZnO interfacial interaction, increasing the lifetime of the quantum dot film to 19.5 ns (Fig. 3d). These results indicate that the PMMA layers help to maintain charge neutrality of quantum dot emitters and preserve their superior emissive properties.

Finally, the strategy of inserting insulating layers between oxide ETLs and quantum dot layers can be extended to improve hybrid QLEDs with other types of quantum dot emitters. Wurtzite-structured CdSe–CdS core–shell nanocrystals with four monolayers of CdS shell<sup>28</sup> are used as an example. When these thin-shell quantum dots are used as emitters, a device structure of ITO/PEDOT:PSS/poly-TPD/PVK/QDs/ZnO/Ag yields a peak EQE of 2.5%. When a 6 nm PMMA layer is inserted, the peak EQE of the QLEDs increases to 4.7% (Extended Data Fig. 8).

The present work demonstrates the best-performing solution-processed red LEDs with colour-saturated emission, record efficiency, low efficiency roll-off, sub-bandgap turn-on voltage, excellent reproducibility and outstanding operational stability, whose overall performance is comparable to state-of-the-art OLEDs produced by vacuum deposition<sup>6–8</sup> (Extended Data Table 1). Such outstanding optoelectronic performance is achieved by introducing the conceptually new device structure resulting from the insertion of an insulating layer between the quantum dot layer and the oxide ETL, and by using quantum dots with superior properties as solution-processed inorganic emissive centres. We believe that there are no fundamental obstacles to extending these techniques to differently coloured QLEDs, which would lead to low-cost, large-area, high-efficiency, high-colour-quality, stable, all-solution-processed electroluminescent devices for both display and solid-state lighting technologies.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 9 May; accepted 29 August 2014.**

**Published online 29 October 2014.**

1. Friend, R. H. *et al.* Electroluminescence in conjugated polymers. *Nature* **397**, 121–128 (1999).
2. Supran, G. J. *et al.* QLEDs for displays and solid-state lighting. *MRS Bull.* **38**, 703–711 (2013).
3. Mashford, B. S. *et al.* High-efficiency quantum-dot light-emitting devices with enhanced charge injection. *Nature Photon.* **7**, 407–412 (2013).
4. Chien, C.-H. *et al.* Electrophosphorescent polyfluorenes containing osmium complexes in the conjugated backbone. *Adv. Funct. Mater.* **18**, 1430–1439 (2008).
5. Yook, K. S. & Lee, J. Y. Small molecule host materials for solution processed phosphorescent organic light-emitting diodes. *Adv. Mater.* **26**, 4218–4233 (2014).
6. Meerheim, R. *et al.* Influence of charge balance and exciton distribution on efficiency and lifetime of phosphorescent organic light-emitting devices. *J. Appl. Phys.* **104**, 014510 (2008).
7. Nakanotani, H. *et al.* High-efficiency organic light-emitting diodes with fluorescent emitters. *Nature Commun.* **5**, 4016 (2014).
8. Murawski, C., Leo, K. & Gather, M. C. Efficiency roll-off in organic light-emitting diodes. *Adv. Mater.* **25**, 6801–6827 (2013).
9. Brus, L. E. Electron-electron and electron-hole interactions in small semiconductor crystallites: the size dependence of the lowest excited electronic state. *J. Chem. Phys.* **80**, 4403–4409 (1984).

10. Murray, C. B., Kagan, C. R. & Bawendi, M. G. Synthesis and characterization of monodisperse nanocrystals and close-packed nanocrystal assemblies. *Annu. Rev. Mater. Sci.* **30**, 545–610 (2000).
11. Peng, X. G. An essay on synthetic chemistry of colloidal nanocrystals. *Nano Res.* **2**, 425–447 (2009).
12. Colvin, V. L., Schlamp, M. C. & Alivisatos, A. P. Light-emitting diodes made from cadmium selenide nanocrystals and a semiconducting polymer. *Nature* **370**, 354–357 (1994).
13. Coe, S., Woo, W.-K., Bawendi, M. G. & Bulovic, V. Electroluminescence from single monolayers of nanocrystals in molecular organic devices. *Nature* **420**, 800–803 (2002).
14. Zhao, J. *et al.* Efficient CdSe/CdS quantum dot light-emitting diodes using a thermally polymerized hole transport layer. *Nano Lett.* **6**, 463–467 (2006).
15. Caruge, J. M., Halpert, J. E., Wood, V., Bulovic, V. & Bawendi, M. G. Colloidal quantum-dot light-emitting diodes with metal-oxide charge transport layers. *Nature Photon.* **2**, 247–250 (2008).
16. Cho, K.-S. *et al.* High-performance crosslinked colloidal quantum-dot light-emitting diodes. *Nature Photon.* **3**, 341–345 (2009).
17. Qian, L., Zheng, Y., Xue, J. & Holloway, P. H. Stable and efficient quantum-dot light-emitting diodes based on solution-processed multilayer structures. *Nature Photon.* **5**, 543–548 (2011).
18. Kwak, J. *et al.* Bright and efficient full-color colloidal quantum dot light-emitting diodes using an inverted device structure. *Nano Lett.* **12**, 2362–2366 (2012).
19. Bae, W. K. *et al.* Controlling the influence of Auger recombination on the performance of quantum-dot light-emitting diodes. *Nature Commun.* **4**, 2661 (2013).
20. Lee, K.-H. *et al.* Over 40 cd/A efficient green quantum dot electroluminescent device comprising uniquely large-sized quantum dots. *ACS Nano* **8**, 4893–4901 (2014).
21. Thesen, M. W. *et al.* Hole-transporting host-polymer series consisting of triphenylamine basic structures for phosphorescent polymer light-emitting diodes. *J. Polym. Sci. A* **48**, 3417–3430 (2010).
22. Lee, D.-H., Liu, Y.-P., Lee, K.-H., Chae, H. & Cho, S. M. Effect of hole transporting materials in phosphorescent white polymer light-emitting diodes. *Org. Electron.* **11**, 427–433 (2010).
23. Sayyah, S. M., Khaliel, A. B. & Moustafa, H. Electronic structure and ground state properties of PMMA polymer: I. Step-by-step formation and stereo-regularity of the polymeric chain—AM1-MO treatment. *Int. J. Polym. Mater.* **54**, 505–518 (2005).
24. Qin, H. Y. *et al.* Single-dot spectroscopy of zinc-blende CdSe/CdS core/shell nanocrystals: nonblinking and correlation with ensemble measurements. *J. Am. Chem. Soc.* **136**, 179–187 (2014).
25. Wellmann, P. *et al.* High-efficiency p-i-n organic light-emitting diodes with long lifetime. *J. Soc. Inf. Disp.* **13**, 393–397 (2005).
26. Greenham, N. C., Friend, R. H. & Bradley, D. D. C. Angular dependence of the emission from a conjugated polymer light-emitting diode: implications for efficiency calculations. *Adv. Mater.* **6**, 491–494 (1994).
27. Javaux, C. *et al.* Thermal activation of non-radiative Auger recombination in charged colloidal nanocrystals. *Nature Nanotechnol.* **8**, 206–212 (2013).
28. Nan, W. N. *et al.* Crystal structure control of zinc-blende CdSe/CdS core/shell nanocrystals: synthesis and structure-dependent optical properties. *J. Am. Chem. Soc.* **134**, 19685–19693 (2012).

**Acknowledgements** This work is financially supported by the National High Technology Research and Development Program of China (2011AA050520), the National Natural Science Foundation of China (21233005 and 51172203), the National Science Funds for Distinguished Young Scholar of Zhejiang Province (R4110189), the Public Welfare Project of Zhejiang Province (2013C31057), the Jiangsu Natural Science Foundation (BK20130006 and BK20131413), the National Basic Research Program of China (2015CB932200) and the Jiangsu Specially-Appointed Professor programme. We thank L. Liao and L. Zhang for assistance in cross-measuring the QLED and OLED devices. We thank Q. Chen for assistance with atomic force microscopy and scanning Kelvin probe microscopy measurements. We thank Z. Zhang and C. Jin for assistance with cross-sectional transmission electron microscopy experiments. We also thank J. Yu and G. Qian for assistance in obtaining the confocal images.

**Author Contributions** Y.J. and X.P. had the idea for and designed the experiments and supervised the work. X.D. carried out the device fabrication and characterizations. Z.Z. conducted the optical measurements and participated in device fabrication. Y.N. and H.C. synthesized the quantum dots. X.L. and L.C. carried out the atomic force microscopy and scanning Kelvin probe microscopy experiments. Y.J. wrote the first draft of the manuscript. X.P. and J.W. provided major revisions. All authors discussed the results and commented on the manuscript.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to Y.J. (yizhengjin@zju.edu.cn) or X.P. (xpeng@zju.edu.cn).

## METHODS

**Materials.** PVK (average molecular weight, 25,000–50,000 g mol<sup>-1</sup>), PMMA (average molecular weight, ~120,000 g mol<sup>-1</sup>) and zinc acetate dihydrate (>98%) were purchased from Sigma Aldrich. Tetramethylammonium hydroxide (TMAH, 98%), 1-dodecanethiol (98%), dodecane (99%), octylamine (98%), cadmium oxide (CdO, 99.998%), 1-octadecene (ODE, 90%) and oleic acid (HOI, 90%) were purchased from Alfa-Aesar. Sodium diethyldithiocarbamate trihydrate (NaDDTC·3H<sub>2</sub>O, 99%) was purchased from Aladdin Reagents. Cadmium acetate dihydrate (Cd(Ac)<sub>2</sub>·2H<sub>2</sub>O, 98.5%) was purchased from Shanghai Tingxin Reagents. Chlorobenzene (extra dry, 99.8%), *m*-xylene (extra dry, 99%), octane (extra dry, >99%), ethanol (extra dry, 99.5%), 2-ethanolamine (99%) and oleylamine (80–90%) were purchased from Acros. Dimethyl sulphoxide (DMSO, HPLC grade) and ethyl acetate (HPLC grade) were purchased from J&K Chemical Ltd. Acetone was purchased from Sinopharm Chemical Reagents. Poly-TPD was purchased from American Dye Source. Patterned ITO-glass substrates (sheet resistance, 15 Ω sq<sup>-1</sup>) were purchased from Xiamen Weihua company. All materials were used as received.

**Synthesis of CdSe–CdS core–shell quantum dots.** The CdSe–CdS core–shell quantum dots (ten monolayers of CdS shell) with phase-pure zinc blende structure were synthesized according to ref. 28 with some modifications. Briefly, the zinc blende CdSe cores (3.1 nm) were synthesized and purified according to our recent report<sup>28</sup>. For the shell coating, dodecane (3.8 ml), octylamine (1.05 ml), oleylamine (0.45 ml) and purified CdSe core solution containing  $2 \times 10^{-7}$  mol of nanocrystals were mixed and heated to 80 °C under an argon flow. Reaction cycles, that is, addition of the Cd(DDTC)<sub>2</sub>-amine precursor solutions at 80 °C and growth of CdS monolayers at 150 °C for ~20 min, were performed for the growth of the first six monolayers. Desirable amounts of Cd(DDTC)<sub>2</sub>-amine solutions, that is, 0.08, 0.12, 0.16, 0.21, 0.26 and 0.32 ml, were used for the growth of first, second, third, fourth, fifth and sixth monolayers, respectively. For the seventh to tenth monolayers of CdS, the precursor solution was changed to 50 mol% of Cd(DDTC)<sub>2</sub> and 50 mol% of Cd(Ol)<sub>2</sub>, and the growth temperature was set at 160 °C. Quantities 0.39, 0.47, 0.55, and 0.64 ml of the precursor solutions were used for the growth of seventh, eighth, ninth and tenth monolayers of CdS, respectively. The resulting CdSe–CdS core–shell nanocrystals were purified and subjected to a ligand exchange procedure. The original ligands were replaced by 1-dodecanethiol. The ligand-exchanged quantum dots were dispersed in octane and filtered before use.

**Colloidal ZnO nanocrystals.** Colloidal ZnO nanocrystals were synthesized by a low-temperature solution-precipitation method<sup>29</sup> with some modifications. A DMSO solution (30 ml) of zinc acetate hydrate (3 mmol) was mixed with an ethanol solution (10 ml) of TMAH (5.5 mmol) and stirred for 24 h under ambient conditions. Then the ZnO nanocrystals were precipitated by adding ethyl acetate and redispersed in ethanol. Additional ligands of 2-ethanolamine (160 μl) was introduced to stabilize the nanoparticles. The ZnO nanocrystals were further washed with ethyl acetate and redispersed in ethanol. The solutions were filtered before use.

**Device fabrication.** PEDOT:PSS solutions (Baytron P VP Al 4083, filtered through a 0.45 μm N66 filter) were spin-coated onto the ITO-coated glass substrates at 4,000 r.p.m. for 60 s and baked at 140 °C for 10 min. The PEDOT:PSS-coated substrates were transferred into a nitrogen-filled glove box (O<sub>2</sub> < 1 p.p.m., H<sub>2</sub>O < 1 p.p.m.). Poly-TPD (in chlorobenzene, 8 mg ml<sup>-1</sup>), PVK (in *m*-xylene, 1.5 mg ml<sup>-1</sup>), quantum dots (in octane, 15 mg ml<sup>-1</sup>), PMMA (in acetone, 1.8 mg ml<sup>-1</sup>) and ZnO nanocrystals (in ethanol, 50 mg ml<sup>-1</sup>) were deposited layer by layer by spin coating at 2,000 r.p.m. for 45 s. The poly-TPD and PVK layers were baked at 110 °C for 20 min and at 170 °C for 30 min, respectively, before the deposition of the next layer. Finally, Ag electrodes (100 nm) were deposited using a thermal evaporation system through a shadow mask under a high vacuum of ~6 × 10<sup>-7</sup> torr. The device area was 4 mm<sup>2</sup> as defined by the overlapping area of the ITO and Ag electrodes. The devices were encapsulated in the glove-box by the cover glasses using ultraviolet-curable resin.

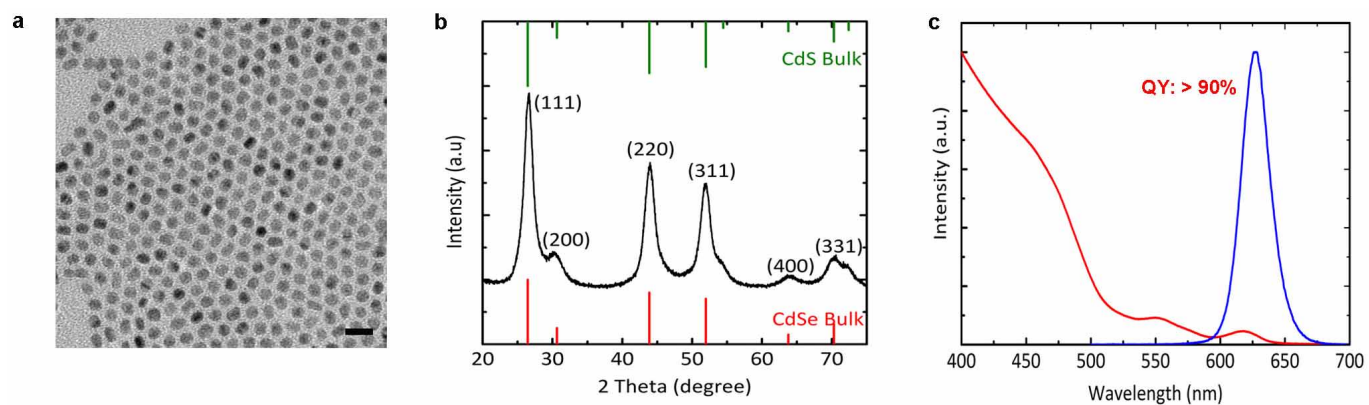
**Characterizations.** We used a Keithley 2400 electrometer for current density–voltage characterizations and a fibre integration sphere (FOIS-1) coupled with a QE-65000 spectrometer for light output measurements (Extended Data Fig. 9). Note that the ITO glass substrates are in close contact with the input port of the integration sphere (but are not inserted into the integration sphere). The area of the QLED device (4 mm<sup>2</sup>) is much smaller than that of the input port (9.5 mm in diameter) so that the coupling factor for the photons emitted into the forward viewing directions<sup>30</sup> (from the QLED to the integration sphere) is unity. The electroluminescence characteristics of the QLEDs were cross-checked using a system comprising a photometer (Spectra Scan PR655) coupled with a computer-controlled Keithley 2400 electrometer in L. Liao's laboratory. The half-lifetime  $T_{50}$  and the driving voltages are measured at the same time using an ageing system made by Shanghai University.

The absorption spectra of the nanocrystals were measured using a Shimadzu UV 3600 spectrophotometer. The photoluminance spectra of the quantum dots were obtained by using an Edinburgh Instruments FLS920 spectrometer. The absolute photoluminance quantum yield of the quantum dot solution was measured using an Ocean Optics FOIS-1 integrating sphere coupled with a QE65000 spectrometer. The time-resolved fluorescence spectra of the quantum dot films were measured by the time-correlated single-photon counting method using an Edinburgh Instruments FLS920 fluorescence spectrometer. The samples were excited by a 405 nm pulsed diode laser (EPL-405). An Olympus confocal laser scanning microscope (FV1000) equipped with an inverted fluorescence microscope (IX81) was used to evaluate the emissive features of the quantum dot films. A 488 nm laser was used to excite the sample, and the signals in the wavelength range of 580–680 nm were collected for imaging. Fourier transform infrared spectroscopy (FTIR) spectra were obtained using a Bruker Vector 27 spectrophotometer.

Transmission electron microscope (TEM) analyses on the cross-sections of the QLEDs were carried out using a Tecnai G2 F20 microscope. The cross-sectional samples were prepared by using focused-ion-beam equipment (Quata 3D FEG). Atomic force microscopy (AFM) measurements were conducted on either a Park XE-120 atomic force microscope or an Agilent 5500 AFM (Agilent Technologies) using silicon AFM tips (HQ:NSC18 and HQ:NSC15, Mikromasch) or high-resolution probes (Hi\*RES-C19/AIBS). Scanning Kelvin probe microscopy measurements were performed on a Park XE-120 atomic force microscope using Cr–Au-coated conducting AFM tips (HQ:NSC19, Mikromasch). The thicknesses of the multilayers were measured using a Dektak 150 stylus profilometer.

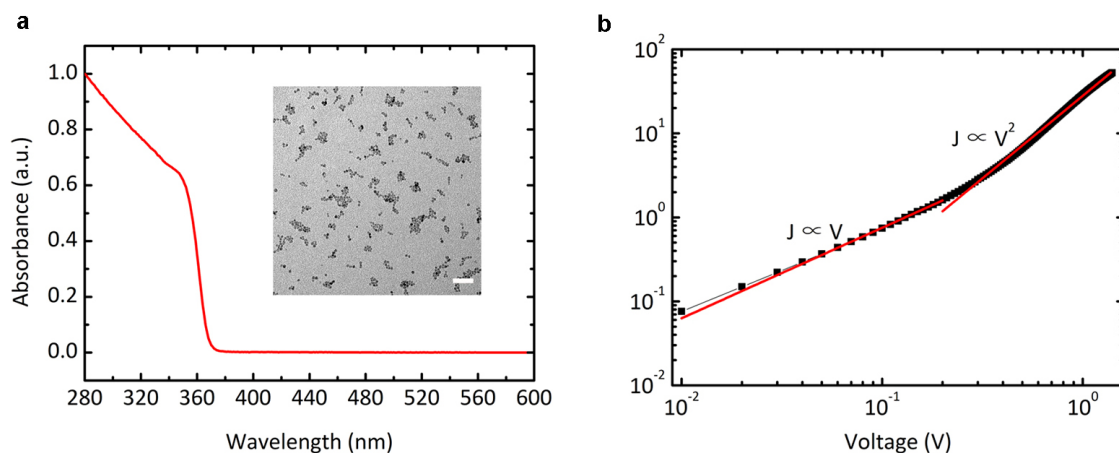
The transmission electron microscopy images of the quantum dots were made on a Hitachi 7700 TEM operated at 80 keV. The X-ray diffraction measurements were conducted using a Rigaku Ultimate-IV system operated at 40 kV and 40 mA using the Cu Kα line ( $\lambda = 1.5418$  Å).

29. Qian, L. *et al.* Electroluminescence from light-emitting polymer/ZnO nanoparticle heterojunctions at sub-bandgap voltages. *Nano Today* **5**, 384–389 (2010).
30. Forrest, S. R., Bradley, D. D. C. & Thompson, M. E. Measuring the efficiency of organic light-emitting devices. *Adv. Mater.* **15**, 1043–1048 (2003).
31. Humphries, M. J., Wilson, R. J., Fernandez, O. & Archer, R. A. Developments in solution processable polymer light-emitting diodes. *J. Photon. Energy* **1**, 011019 (2011).
32. Giridhar, T. *et al.* An electron transporting unit linked multifunctional Ir(III) complex: a promising strategy to improve the performance of solution-processed phosphorescent organic light-emitting diodes. *Chem. Commun.* **50**, 4000–4002 (2014).
33. Andrade, B. D. *et al.* Phosphorescent OLEDs with saturated colors. *Proc. SPIE* **6655**, 66550G (2007).
34. Uoyama, H., Goushi, K., Shizu, K., Nomura, H. & Adachi, C. Highly efficient organic light-emitting diodes from delayed fluorescence. *Nature* **492**, 234–238 (2012).



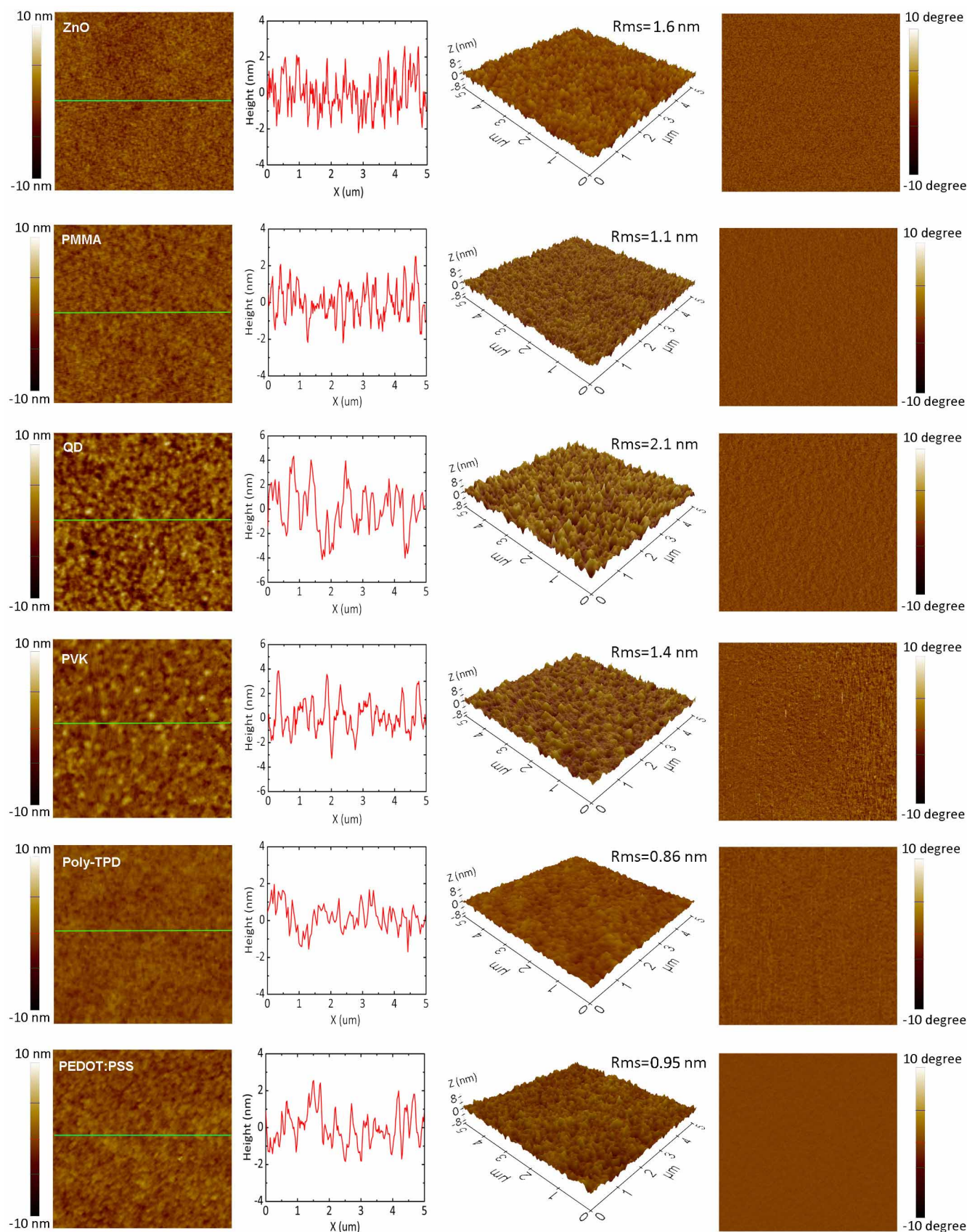
**Extended Data Figure 1 | CdSe-CdS core-shell quantum dots with ten monolayers of CdS shell.** **a**, A typical TEM image. Scale bar, 20 nm. **b**, X-ray diffraction profile. **c**, Ultraviolet-visible absorption and photoluminescence spectra.





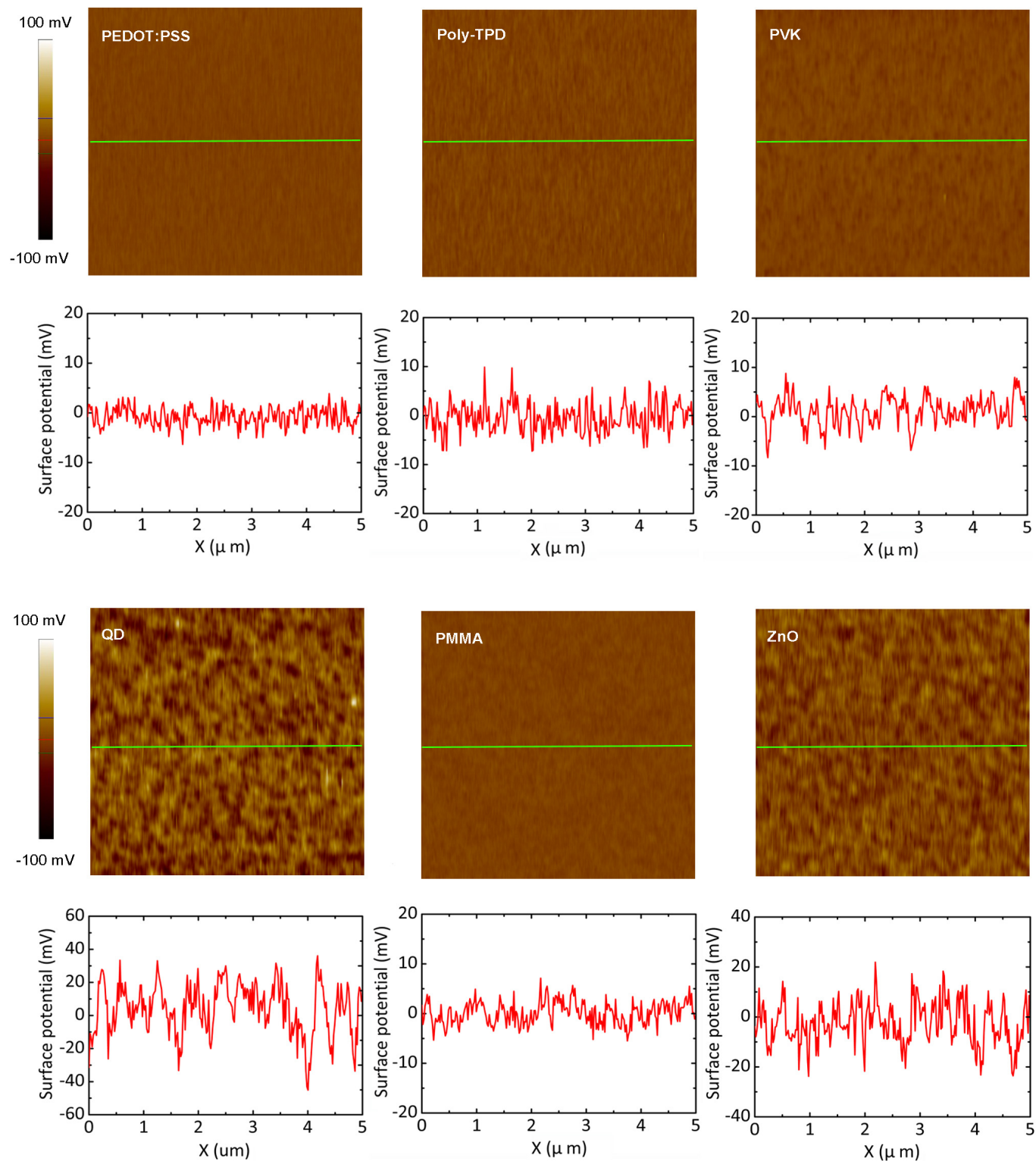
**Extended Data Figure 2 | ZnO nanocrystals as ETLs.** **a**, Ultraviolet–visible absorption spectrum and a typical TEM image (inset: scale bar, 50 nm) of the colloidal ZnO nanocrystals. **b**, Current density–voltage ( $J$ - $V$ ) characteristics of an electron-only device (ITO/Al/ZnO/Al). The thickness of the ZnO layer is 300 nm. The electron mobility of the ZnO film is obtained by fitting

space-charge-limited-current region ( $J \propto V^2$ ) with Child's law,  $J = (9/8)\epsilon_r\epsilon_0\mu_e V^2/d^3$ , where  $\epsilon_0$ ,  $\epsilon_r$ ,  $\mu_e$  and  $d$  are the vacuum permittivity, relative permittivity, electron mobility and film thickness, respectively<sup>17</sup>. By assuming that  $\epsilon_r = 4$ ,  $\mu_e$  is determined to be  $1.8 \times 10^{-3} \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$ .



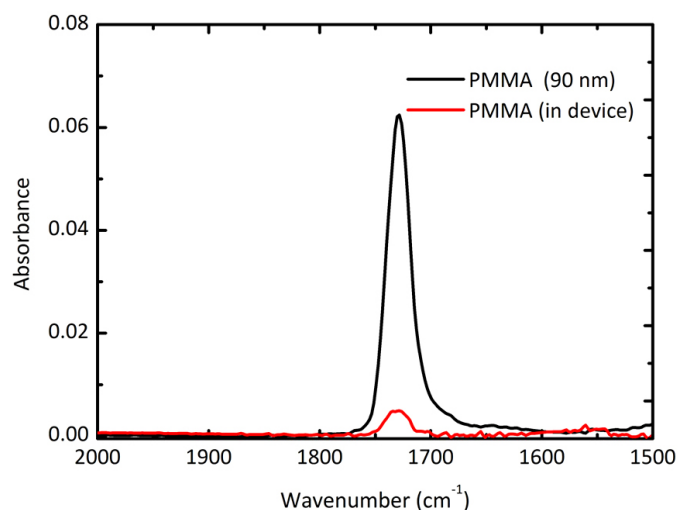
**Extended Data Figure 3 | AFM characterizations of the multilayers of PEDOT:PSS, poly-TPD, PVK, quantum dots, PMMA and ZnO films in the device configuration, respectively.** For each layer, the height image, the line-scan profile, the pseudo-three-dimensional image and the phase image are shown. Note that the surface root mean squared values may change owing

to the tip-to-tip and sample-to-sample variations. Extensive AFM measurements show that the root mean squared roughnesses for the quantum dot layer and the PMMA layer are in the ranges of 1.6–2.6 nm and 0.6–1.6 nm, respectively.

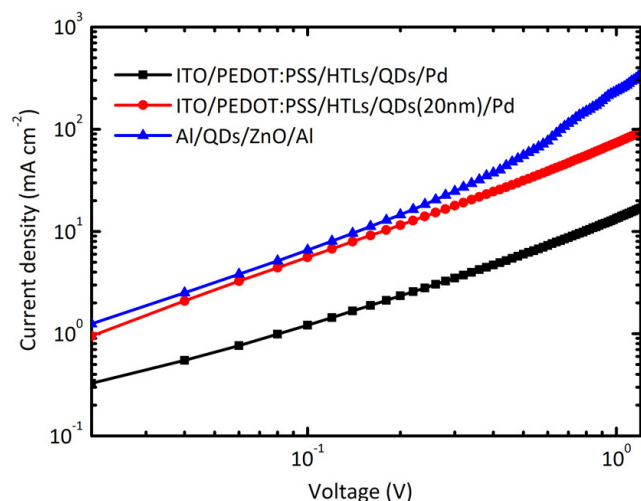


**Extended Data Figure 4 | Scanning Kelvin probe microscopy characterizations of the multilayers of PEDOT:PSS, poly-TPD, PVK, quantum dots, PMMA and ZnO films in the device configuration.** Note that the data have been linearly fitted to show the spatial uniformity.

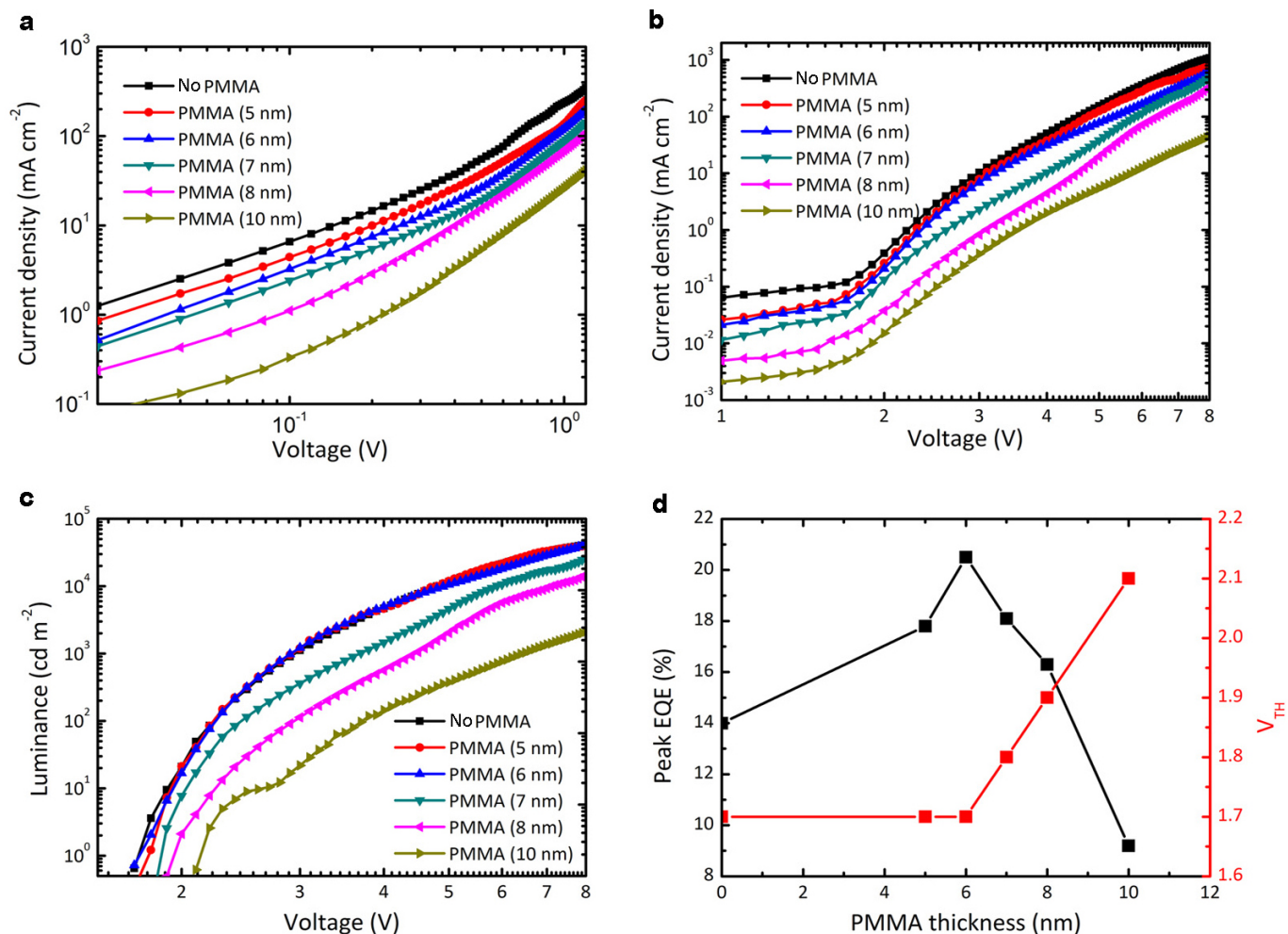




**Extended Data Figure 5 | FTIR analyses to determine the average thickness of the PMMA layers used in the QLEDs.** The sample for FTIR measurements was produced by layer-by-layer spin-coating the PVK (in *m*-xylene,  $1.5 \text{ mg ml}^{-1}$ ), quantum dots (in octane,  $15 \text{ mg ml}^{-1}$ ) and PMMA (in acetone,  $1.8 \text{ mg ml}^{-1}$ ) at 2,000 r.p.m. onto the cleaned  $\text{CaF}_2$  substrates. We assume that the average thickness of the PMMA layer in this sample is identical to that of the PMMA layer in the optimized QLEDs. The absorption of the carbonylic groups of this sample was measured multiple times, averaged and compared with that of a 90 nm PMMA film (determined by stylus profilometer). Given that the absorbance of the carbonylic groups is in the dynamic range of the instrument, the thickness of the PMMA layer deposited onto the quantum dot film was determined to be  $\sim 6 \text{ nm}$ .



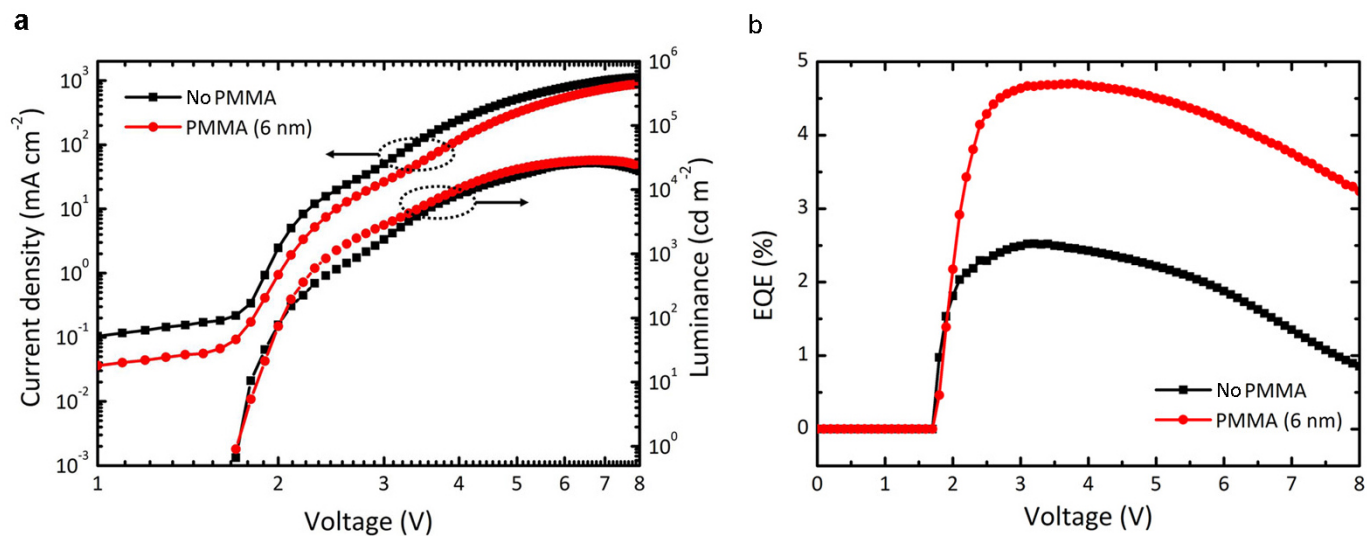
**Extended Data Figure 6 | Electrical measurements on the electron-only devices (ITO/Al/QDs/ZnO/Al) and the hole-only devices (ITO/PEDOT:PSS/poly-TPD/PVK/QDs/Pd).** The current density of the electron-only device (ITO/Al/QDs/ZnO/Al) is more than one order of magnitude greater than that of the hole-only device (ITO/PEDOT:PSS/poly-TPD/PVK/QDs/Pd). In the above two devices, the thicknesses of all layers are identical to those used in the QLEDs. For the quantum dot layer, the thickness is  $\sim 40$  nm. We note that for quasi-type-II CdSe–CdS quantum dots, the electron wavefunction extends to the shell region, whereas the hole wavefunction remains confined to the CdSe core, leading to greater electron mobility than hole mobility. We presume that the recombination zone is close to the PVK/QDs interface due to the very low hole mobility of the quantum dot films. Therefore we also fabricated a hole-only device with a quantum dot layer of  $\sim 20$  nm. The results show that the current density of the hole-only device with the 20 nm quantum dot layer is still much smaller than that of the electron-only device with the 40 nm quantum dot layer.



**Extended Data Figure 7 | Impact of the thickness of the PMMA layer on the QLED performance.** **a**, Current density–applied bias curves for the electron-only devices (ITO/Al/QDs/PMMA/ZnO/Al). **b**, **c**, Current density–driving voltage (b) and luminance–driving voltage (c) curves for the QLEDs (ITO/PEDOT:PSS/poly-TPD/PVK/QDs/PMMA/ZnO/Ag). **d**, Dependence

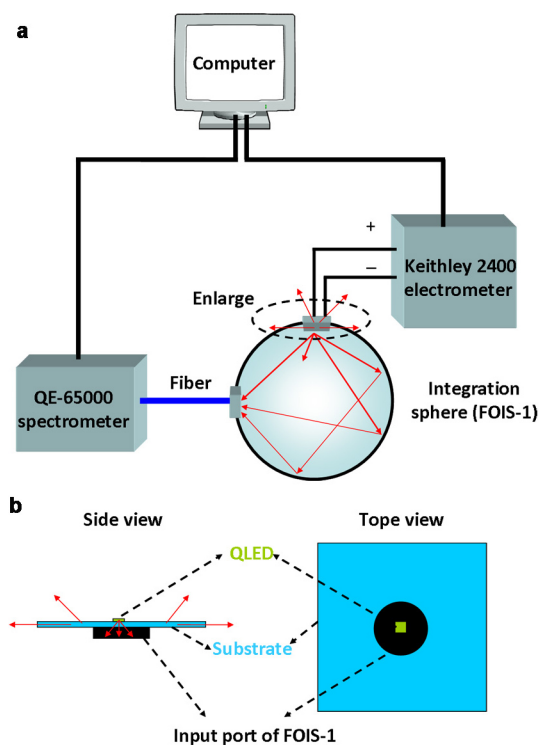
of peak EQEs and turn-on voltages ( $V_{\text{th}}$ ) of the QLEDs on the thicknesses of the PMMA layers. The PMMA layers with thicknesses of 5, 6, 7, 8 and 10 nm were deposited from acetone solutions with concentrations of 1.5, 1.8, 2.1, 2.4 and  $3.0 \text{ mg ml}^{-1}$ , respectively.





**Extended Data Figure 8 | QLEDs with four-monolayer-shell CdSe-CdS quantum dots as emitters. a,** Current density and luminance versus driving voltage characteristics. **b,** Curves of EQE versus driving voltage. The thicknesses

of the PEDOT:PSS, poly-TPD, PVK, quantum dot, PMMA and ZnO layers are 35, 30, 5, 30, 6 and 150 nm, respectively.



**Extended Data Figure 9 | QLED characterization system.** **a**, A Keithley 2400 electrometer is used to obtain current density–voltage characteristics. A fibre integration sphere (FOIS-1) coupled with a QE-65000 spectrometer is used for light output measurements. **b**, Top-view and side-view of the QLEDs in direct contact with the input port of the fibre integration sphere. The glass substrate ( $19\text{ mm} \times 19\text{ mm}$ ) is rested on top of (but not inserted into) the integration sphere. The area of the QLED device ( $4\text{ mm}^2$ ) is much smaller than that of the input port ( $9.5\text{ mm}$  in diameter) of the integration sphere so that the coupling factor for the photons emitted into the forward viewing directions (from the QLED to the integration sphere) is unity.

Extended Data Table 1 | Comparison of our device with other high-performance red LEDs

Device <sup>*</sup>	Peak EQE (%)	EQE(%) @ 100 mA cm <sup>-2</sup>	<sup>†</sup> Wall-plug E <sub>eff</sub> (%) @ 100 mA cm <sup>-2</sup>	V <sub>th</sub> (V)	CIE coordinates	<sup>‡</sup> T <sub>50</sub> @ 100 cd m <sup>-2</sup>
Our QLED	20.5	15.1	5.5	1.7	(0.71,0.29)	100,000
QLED <sup>§</sup>	18.5	12	4.8	1.5	N/A	4,000
PLED <sup>¶</sup>	18	15	2.5	6.0	(0.64,0.35)	N/A
PLED <sup>§1</sup>	13	N/A	N/A	N/A	(0.63,0.32)	N/A
S-OLED <sup>§2</sup>	20.59	N/A	N/A	6.0	(0.67,0.33)	N/A
Ph-OLED <sup>§</sup>	20	<10	<5.3	2.4	(0.63,0.37)	> 1000,000
Ph-OLED <sup>§3</sup>	17.7	N/A	N/A	N/A	(0.67,0.33)	N/A
TADF-OLED <sup>7, §4</sup>	17.5	11	3.6	3.0	(0.61,0.39)	< 40,000 <sup>§</sup>

Our device is compared with the other high-performance red QLEDs, polymer LEDs (PLEDs), solution-processed small-molecule OLEDs (S-OLEDs) and vacuum-deposited OLEDs including both phosphorescence organic LEDs (Ph-OLEDs) and those using thermally activated delayed fluorescence emitters (TADF-OLEDs) in literature reports. Key parameters including peak EQE, EQE at 100 mA cm<sup>-2</sup>, wall-plug efficiency (wall-plug E<sub>eff</sub>) at 100 mA cm<sup>-2</sup>, turn-on voltage (V<sub>th</sub>), CIE coordinates and T<sub>50</sub> lifetime are listed. The parameters of our device are highlighted as blue. The parameters superior or inferior to those of our devices are highlighted as green or red, respectively.

<sup>\*</sup> This table does not include vacuum-deposited OLEDs with optical engineering (light coupling structures or horizontally aligned emitters) or tandem structures.

<sup>†</sup> The wall-plug E<sub>eff</sub> is estimated from EQE and photon energy at the electroluminescence peak.

<sup>‡</sup> The T<sub>50</sub> lifetime is converted from the values for different initial luminances to the values for an initial luminance of 100 cd m<sup>-2</sup> by assuming an acceleration factor of 1.5.

<sup>§</sup> In ref. 7, the lifetime data are available only for an orange-emitting device.



# Asymmetric photoredox transition-metal catalysis activated by visible light

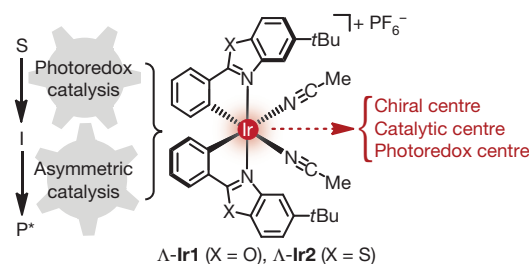
Haohua Huo<sup>1</sup>, Xiaodong Shen<sup>1</sup>, Chuanyong Wang<sup>1</sup>, Lili Zhang<sup>1</sup>, Philipp Röse<sup>1</sup>, Liang-An Chen<sup>2</sup>, Klaus Harms<sup>1</sup>, Michael Marsch<sup>1</sup>, Gerhard Hilt<sup>1</sup> & Eric Meggers<sup>1,2</sup>

Asymmetric catalysis is seen as one of the most economical strategies to satisfy the growing demand for enantiomerically pure small molecules in the fine chemical and pharmaceutical industries<sup>1</sup>. And visible light has been recognized as an environmentally friendly and sustainable form of energy for triggering chemical transformations and catalytic chemical processes<sup>2–5</sup>. For these reasons, visible-light-driven catalytic asymmetric chemistry is a subject of enormous current interest<sup>2–5</sup>. Photoredox catalysis provides the opportunity to generate highly reactive radical ion intermediates with often unusual or unconventional reactivities under surprisingly mild reaction conditions<sup>6</sup>. In such systems, photoactivated sensitizers initiate a single electron transfer from (or to) a closed-shell organic molecule to produce radical cations or radical anions whose reactivities are then exploited for interesting or unusual chemical transformations. However, the high reactivity of photoexcited substrates, intermediate radical ions or radicals, and the low activation barriers for follow-up reactions provide significant hurdles for the development of efficient catalytic photochemical processes that work under stereochemical control and provide chiral molecules in an asymmetric fashion<sup>7</sup>. Here we report a highly efficient asymmetric catalyst that uses visible light for the necessary molecular activation, thereby combining asymmetric catalysis and photocatalysis. We show that a chiral iridium complex can serve as a sensitizer for photoredox catalysis and at the same time provide very effective asymmetric induction for the enantioselective alkylation of 2-acyl imidazoles. This new asymmetric photoredox catalyst, in which the metal centre simultaneously serves as the exclusive source of chirality, the catalytically active Lewis acid centre, and the photoredox centre, offers new opportunities for the ‘green’ synthesis of non-racemic chiral molecules.

Recently, strategies have been developed in which efficient catalytic photochemical processes that work under stereochemical control and provide chiral molecules in an asymmetric fashion can be carried out by two catalysts that work in tandem for a single chemical transformation<sup>8</sup>. In such dual-catalyst reactions, visible-light redox sensitizers are combined with asymmetric co-catalysts, such as chiral secondary amines<sup>9–13</sup>, chiral *N*-heterocyclic carbenes<sup>14</sup>, chiral Brønsted acids<sup>15</sup>, chiral Lewis acids<sup>16</sup>, or chiral thiourea<sup>17</sup>. With respect to single catalysts, ultraviolet light in combination with hydrogen bonding or Lewis acid interaction has been used previously in pioneering work to trigger enantioselective catalysis<sup>18–20</sup>, and an enantioselective cycloaddition induced by visible light—although not including photoinduced electron transfer—has been reported<sup>21</sup>; also, an interesting but special case of photoactivated enamine catalysis was disclosed recently in which a transient electron donor–acceptor complex is capable of absorbing visible light and triggering a charge transfer<sup>22</sup>. General solutions for interfacing visible-light-induced photoredox chemistry and asymmetric catalysis with single catalysts are highly desirable, and will potentially provide new opportunities for reaction design by having a closer control over the entire reaction path, including the crucial stereodiscrimination step.

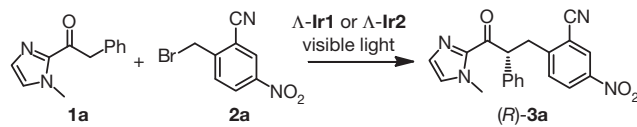
Taking into account that currently used visible-light photosensitizers are typically based on transition-metal complexes<sup>2–5</sup>, and that chiral

transition-metal complexes constitute an established class of catalysts for asymmetric transformations<sup>1</sup>, we envisioned the combination of these two features into a single transition-metal-based asymmetric photoredox catalyst. We conducted our study with the recently developed chiral-at-metal iridium(III) complex  $\Lambda$ -Ir1<sup>23</sup> and the derivative  $\Lambda$ -Ir2 (Fig. 1). In both complexes, the octahedral iridium centre is coordinated by two achiral bidentate ligands in a left ( $\Lambda$ )- or right ( $\Delta$ )-handed propeller-type fashion, thereby establishing metal-centred chirality<sup>24,25</sup>, and coordinated by two additional labile acetonitriles which give access to a Lewis acid metal centre upon ligand exchange (see Supplementary Fig. 1 for a crystal structure of  $\Lambda$ -Ir2). Our laboratory has previously reported the activation of  $\alpha,\beta$ -unsaturated 2-acyl imidazoles by chiral-at-metal  $\Lambda$ - or  $\Delta$ -Ir1 as a step towards the enantioselective addition of indole nucleophiles<sup>23</sup>, so we considered the possibility that these chiral Lewis acids might be capable of intertwining chiral enolate catalysis<sup>26</sup> with photoredox radical ion chemistry<sup>2–5</sup>: we therefore selected the model reaction of 2-acyl imidazole **1a** with the electron deficient benzyl bromide **2a** as our starting point (see Table 1). Encouragingly, in the presence of light from a 14 W energy-saving household lamp,  $\Lambda$ -Ir1 at a loading of 5 mol% was able to catalyse the reaction between **1a** and **2a**, providing the  $\alpha$ -alkylation product **3a** in good yield (85%) and with high enantioselectivity (95% enantiomeric excess, e.e.) after 20 h photolysis at room temperature (entry 1 of Table 1). Optimization of the reaction conditions—by empirically adjusting the solvent, increasing the concentration to speed up the reaction, slightly raising the temperature to promote ligand exchange at the iridium centre, and adding the weak base Na<sub>2</sub>HPO<sub>4</sub> to facilitate enolate chemistry—provided the product **3a** in an excellent yield of 97% with 95% e.e. after exposure to visible light for just 3 h in the presence of a reduced catalyst loading of just 2 mol%  $\Lambda$ -Ir1 (entry 2). The catalyst  $\Lambda$ -Ir2 (2 mol%) even provided the  $\alpha$ -alkylation product in quantitative yield with a superior enantioselectivity of 99% e.e. and a further reduced reaction time of 1.5 h (entry 3). We attribute the improved enantioselectivity to an increased steric hindrance in  $\Lambda$ -Ir2 compared to  $\Lambda$ -Ir1 created by the long C–S bonds of the benzothiazole moieties, which position the two *tert*-butyl groups somewhat closer to the exchange-labile acetonitrile ligands (see Supplementary Fig. 2). The loading of the



**Figure 1 | Chiral iridium complexes for asymmetric photoredox catalysis.** S, substrate; I, intermediate; P\*, non-racemic chiral product. The Ir centre acts as a chiral centre, a catalytic centre, and a photoredox centre.

<sup>1</sup>Fachbereich Chemie, Philipps-Universität Marburg, Hans-Meerwein-Straße, 35043 Marburg, Germany. <sup>2</sup>College of Chemistry and Chemical Engineering, Xiamen University, Xiamen 361005, China.

**Table 1** | Initial iridium-catalysed photoinduced enantioselective alkylation of acyl imidazole **1a** with benzyl bromide **2a**


Entry	Catalyst	Illumination*	Reaction conditions†	t (h)	Yield‡ (%)	e.e.§ (%)
1	Δ-Ir1 (5 mol%)	Visible light	<b>1a</b> (0.3 M, 3 equiv.), MeOH, RT	20	85	95
2	Δ-Ir1 (2 mol%)	Visible light	Na <sub>2</sub> HPO <sub>4</sub> , <b>1a</b> (1.2 M, 3 equiv.), MeOH/THF (4:1), 40 °C	3	97	95
3	Δ-Ir2 (2 mol%)	Visible light	Same as above	1.5	100	99
4	Δ-Ir2 (0.5 mol%)	Visible light	Same as above	4.5	97	98
5	Δ-Ir2 (2 mol%)	Dark	Same as above	1.5	<5	ND
6	None	Visible light	Same as above	16	0	NA

RT, room temperature; ND, not determined; NA, not applicable.

\* Light source: 14 W white light energy-saving lamp.

† All reactions performed under the exclusion of air. See Supplementary Methods for more details.

‡ Isolated yields.

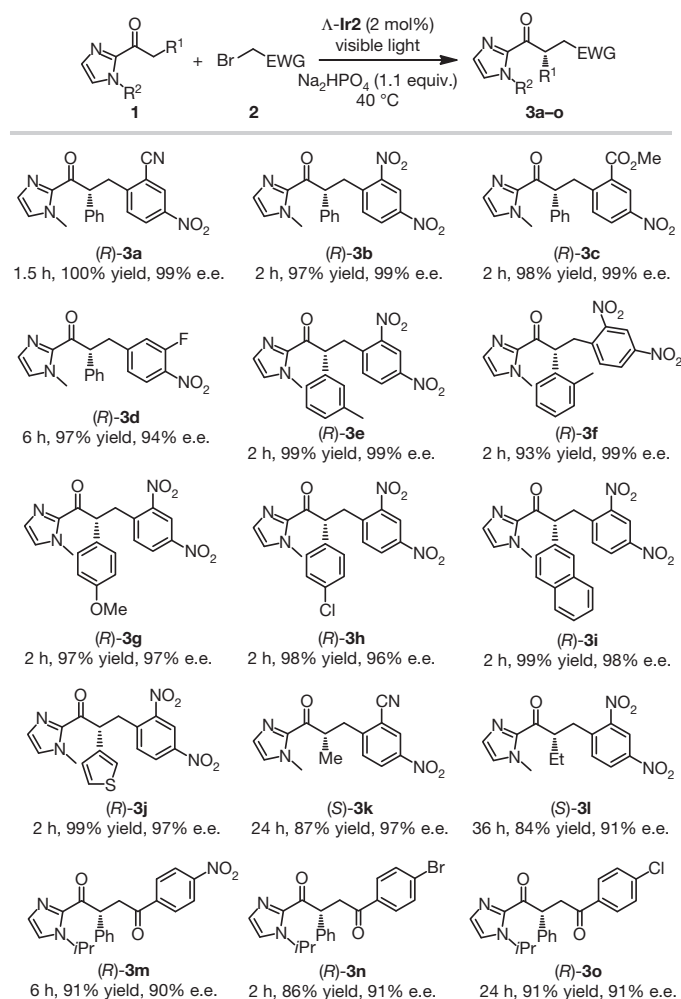
§ Enantiomeric excess determined by HPLC analysis on chiral stationary phase.

catalyst Δ-Ir2 can be further decreased to merely 0.5 mol% without much affecting the yield (97%) or the enantioselectivity (98% e.e.) (entry 4). We note that neither the catalyst Δ-Ir2 alone in the dark (entry 5) nor visible light in the absence of the catalyst (entry 6) trigger this reaction to a significant degree under these conditions, thus unequivocally demonstrating that it is the combination of iridium(III) complex and visible light that is necessary to efficiently catalyse the enantioselective C–C bond formation.

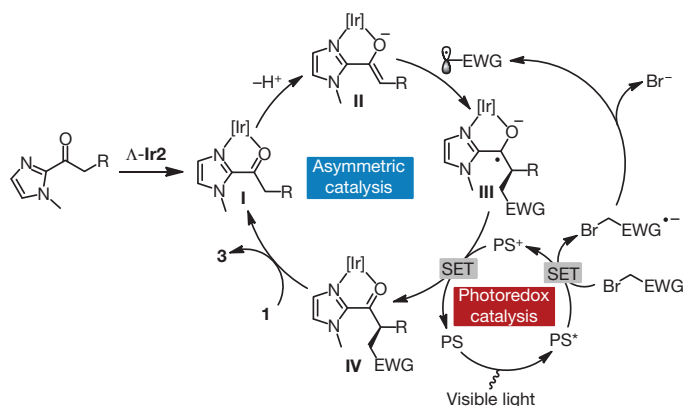
Examples of the photoinduced enantioselective α-alkylation of 2-acyl imidazoles with benzyl bromides catalysed by Δ-Ir2 are summarized in Fig. 2. A variety of electron acceptor substituted benzyl bromides provide the α-alkylation products in up to quantitative yields (97–100%) and with up to almost perfect enantioselectivities (94–99% e.e.), while requiring only short reaction times of 1.5 to 6 h (**3a–d**). The 2-acyl-*N*-methylimidazole substrates tolerate steric (products **3e** and **3f**), electron donating (product **3g**) and electron accepting (product **3h**) substituents in the phenyl moiety, which can be replaced by the bicyclic aromatic naphthalene (product **3i**) or the heteroaromatic thiophene (product **3j**). Furthermore, the photoredox catalysed reaction also tolerates less acidic 2-acyl-*N*-methylimidazoles devoid of any aromatic substituent at the methylene group, as demonstrated for the products **3k** and **3l**. For these substrates, the addition of a weak base is essential to achieve high conversions and excellent enantioselectivities. We also tested a different class of electrophiles, namely phenacyl bromides, and found that they readily provide the expected C–C bond formation products with very good yields of 86–91% and high enantioselectivities of 90–91% e.e. (products **3m–o**). In order to reach satisfactory enantioselectivities, the *N*-methyl substituent at the imidazole moiety needed to be replaced by the more bulky isopropyl group. Overall, it can be concluded that Δ-Ir2 is a highly effective catalyst for the α-alkylation of acyl imidazoles with acceptor substituted benzyl bromides and phenacyl bromides in the presence of visible light with high to quantitative yields and impressive enantioselectivities, while only using a catalyst loading of 2 mol%.

A plausible mechanism in which photoredox catalysis intertwines with asymmetric catalysis is shown in Fig. 3. Herein, the catalysis is initiated by the coordination of 2-acyl imidazoles (**1**) to the iridium catalyst in a bidentate fashion (intermediate **I**), followed by the formation of a nucleophilic iridium(III) enolate complex (intermediate **II**) upon deprotonation. The subsequent chirality generating key step constitutes the exergonic addition of a photo-reductively generated electrophilic radical to the electron rich metal-coordinated enolate double bond, thereby affording an iridium-coordinated ketyl radical (intermediate **III**). Oxidation of this ketyl intermediate to a ketone by single electron transfer regenerates the iridium(III) photosensitizer and provides the iridium-coordinated product (complex **IV**), which is released upon exchange with unreacted starting material, followed by a new catalytic cycle. The proposed key intermediate which uniquely connects the asymmetric catalysis with the photoredox cycle is the iridium(III) enolate complex **II**,

which not only provides the crucial asymmetric induction in the catalysis cycle and but at the same time serves as the *in situ* generated active chiral photosensitizer<sup>27</sup>.



**Figure 2** | Substrate scope of the photoinduced enantioselective alkylation of 2-acyl imidazoles with acceptor substituted benzyl bromides and phenacyl bromides. Top row, the studied reaction; all other rows show products, giving reaction time, isolated yields after chromatographic purification, and enantiomeric excess (e.e.), which was determined by HPLC on a chiral stationary phase. Product (*S*)-**3k**: for comparison, in the absence of base a yield of 18% with 91% e.e. was obtained after photolysis for 24 h. Product (*S*)-**3l**: reaction was irradiated instead with a blue LED light source (3 W) in order to improve the yield. EWG, electron withdrawing group.

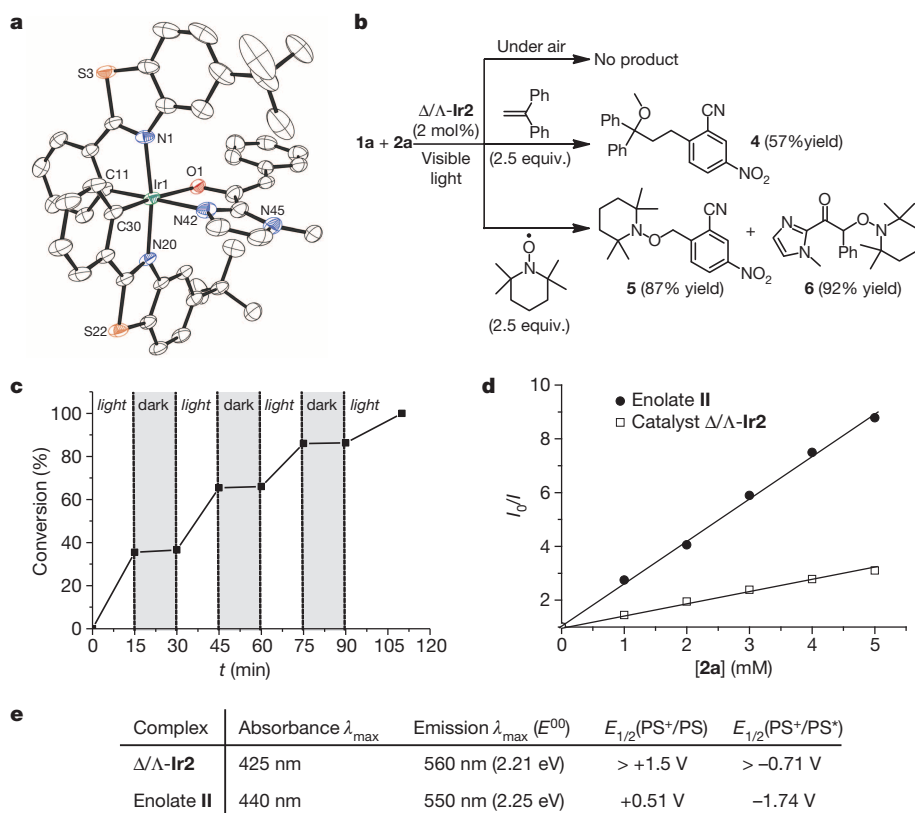


**Figure 3 | Plausible mechanism for a combined photoredox and asymmetric catalysis.** For variations of this mechanism, see Supplementary Fig. 7. SET, single electron transfer; EWG, electron withdrawing group; PS, photosensitizer in the form of enolate complex II. See main text for details.

A series of investigations support this mechanism. To start with, as confirmed by X-ray crystallography (Supplementary Fig. 3), 2-acyl-*N*-methylimidazoles efficiently coordinate to the iridium catalyst  $\Delta/\Lambda$ -Ir2 in a bidentate fashion upon release of the two monodentate acetonitrile ligands, thereby providing the proposed intermediate complex I. Subsequent deprotonation generated the intermediate iridium(III) enolate complex II, which was independently isolated and unambiguously characterized by X-ray crystallography, as shown in Fig. 4a. This structure also illustrates that one face of the prochiral enolate  $\pi$ -bond is blocked by a *tert*-butyl group, thereby rationalizing the observed high asymmetric induction in the course of the proposed diastereoselective addition of the electron-deficient radical to the electron-rich double bond<sup>28</sup>. The determined absolute configurations of the  $\alpha$ -alkylation products are consistent with this mechanistic picture. This reaction step relates to recent reports of the photosensitized generation of electron-deficient radicals

and their stereoselective addition to electron-rich  $\pi$  systems of chiral enamines<sup>9–13</sup>. For our system, a radical mechanism is consistent with the observation that photoreactions in the presence of air, the alkene 1,1-diphenylethylene (isolation of adduct 4) or the radical trap 2,2,6,6-tetramethylpiperidine-1-oxyl (TEMPO) (isolation of products 5 and 6) suppress the formation of the  $\alpha$ -alkylation product (Fig. 4b). Furthermore, the direct correlation between photolysis and product formation is demonstrated by a light–dark interval reaction shown in Fig. 4c.

Importantly, several experimental results lead to the conclusion that the intermediate iridium(III) enolate complex II is not only a key nucleophilic intermediate in the asymmetric catalysis cycle, but also constitutes the active *in situ* assembled photosensitizer in the photoredox cycle. Since iridium(III) complex II apparently represents the only neutral iridium(III) complex within the reaction mixture, this conclusion is also consistent with observed trends regarding redox and photophysical properties of iridium(III) complexes—namely that neutral *bis*-cyclometalated iridium(III) photosensitizers are significantly stronger photoreducing agents than their cationic counterparts<sup>4</sup>. Accordingly, Stern–Volmer plots (Fig. 4d) illustrate that the luminescence emission of the enolate complex II is quenched by benzyl bromide 2a much more efficiently compared to Ir2, which can be attributed to a fast electron transfer from a triplet excited state of enolate complex II to the electron-deficient benzyl bromide. This is furthermore supported by cyclic voltammetry (Supplementary Fig. 4), which reveals that the enolate complex II has a significantly decreased oxidation potential (by around 1 V) compared to the cationic complex Ir2, and thus comprises a much stronger reducing agent in the ground state and even more so in its photoexcited state (Fig. 4e). The estimated excited state redox potential  $E_{1/2}(\text{II}^+/\text{II}^*)$  of  $-1.74$  V versus Ag/AgCl for the enolate complex II is comparable to that of *fac*-[Ir(ppy)<sub>3</sub>] (ppy = 2-phenylpyridine)<sup>4</sup>, an iridium sensitizer that has been used previously for the reductive cleavage of electron deficient benzyl bromides<sup>10</sup>. Conveniently, compared to Ir2 and established iridium(III) photosensitizers<sup>4</sup>, the enolate complex II displays a bathochromically shifted long wavelength absorbance maximum with an additional shoulder at around 500 nm, thus permitting an excitation



**Figure 4 | Mechanistic investigations.** **a**, X-ray crystal structure of the proposed Ir(III) enolate complex intermediate II. This compound was crystallized as a racemic mixture, and only the  $\Lambda$ -enantiomer is shown here, as an ORTEP drawing with 30% probability ellipsoids. **b**, Control experiments in the presence of molecules which react with radicals. See main text for details. **c**, Light–dark interval experiment for the reaction **1a** + **2a**  $\rightarrow$  **3a** according to entry 3 of Table 1. **d**, Luminescence quenching experiments.  $I_0$  and  $I$  are respectively luminescence intensities in the absence and presence of the indicated concentrations of the electron deficient benzyl bromide **2a**. **e**, Comparison of the photo and redox properties of catalyst  $\Delta/\Lambda$ -Ir2 and the enolate complex II. Columns 2 and 3 show wavelength of maximum absorbance and emission, respectively.  $E^{00}$ , energy of the emitting excited state as calculated from the luminescence peak. Column 4 shows the one-electron redox potential of the couple oxidized sensitizer / sensitizer as determined from the peak maximum of differential pulse voltammetry. Column 5 shows the one-electron redox potential of the couple oxidized sensitizer / excited sensitizer as calculated from  $E_{1/2}(\text{PS}^+/\text{PS}^*) = E_{1/2}(\text{PS}^+/\text{PS}) - E^{00}$ .



across half of the visible spectrum ranging from violet to green light (Fig. 4e and Supplementary Fig. 5). The visible-light absorbance of enolate complex **II** is not affected by the presence of organic bromide substrates, thus most probably ruling out the possibility that an electron donor–acceptor complex between enolate complex **II** and bromide substrate is responsible for the light absorption (Supplementary Fig. 6)<sup>22</sup>. It is also worth noting that an independently synthesized enolate complex **II** is catalytically competent and catalyses the photoredox reaction with an identical efficiency compared to **Ir2**, thereby supporting the notion that complex **II** has a dual function as a chiral nucleophile in the catalytic cycle and the *in situ* photosensitizer in the photoredox cycle.

Finally, two distinct variations of the outlined mechanism need to be considered (Supplementary Fig. 7). First, instead of regenerating the photooxidized sensitizer in every cycle ( $\text{PS}^+ + \text{e}^- \rightarrow \text{PS}$ , Fig. 3), the ketyl intermediate **III** might transfer a single electron directly to another bromide substrate, thus skipping the photoredox cycle and leading to a chain reaction. Although the direct light-dependence of the asymmetric photoactivated catalysis shown in Fig. 4c suggests that the asymmetric catalysis and photoredox cycle operate in concert at least to some extent, a contribution of the chain propagation mechanism as a function of the nature of the substrates is feasible, and might even explain the differences in photolysis times for the individual reactions<sup>29</sup>. The second mechanistic variation to discuss revolves around the direct reaction of the intermediate benzyl radical with the oxidized sensitizer ( $\text{PS}^+$ ). However, a major contribution of this recombination process is unlikely since both reactive intermediates will not be generated in close proximity, considering that the fragmentation of the formed radical anion does not occur instantaneously. In this respect, it has been established that the life time of such radical anions significantly increases in protic solvents and with a decreasing energy of the  $\pi^*$  orbitals<sup>30</sup>. This notion is supported by an experiment in which low concentrations of the radical trap TEMPO were still able to capture the intermediate benzyl radical (see Supplementary Methods), thus rendering unlikely an efficient recombination of the benzyl radical with the oxidized iridium sensitizer, and instead favouring an addition of the intermediate electron-deficient benzyl radical with the electron-rich  $\pi$ -bond of the iridium(III) enolate complex **II**, which is present in solution at a much higher steady state concentration.

We have reported a unique case of visible-light-induced asymmetric redox catalysis by a single, structurally simple catalyst. The two catalytic cycles are apparently connected through an intermediate iridium(III) enolate complex, formed from the initial catalyst and the 2-acyl imidazole substrate, which is not only the key nucleophilic intermediate in the asymmetric catalysis cycle but also constitutes the *in situ* generated active visible-light photosensitizer. The reaction scheme that we introduce here may serve as a blueprint for the design of other catalytic asymmetric photoredox reactions, and will most probably provide new avenues for the efficient and green synthesis of non-racemic chiral molecules.

Received 4 August; accepted 23 September 2014.

- Walsh, P. J. & Kozlowski, M. C. *Fundamentals of Asymmetric Catalysis* (University Science Books, 2009).
- Zeitler, K. Photoredox catalysis with visible light. *Angew. Chem. Int. Edn* **48**, 9785–9789 (2009).
- Narayanan, J. M. R. & Stephenson, C. R. J. Visible light photoredox catalysis: applications in organic synthesis. *Chem. Soc. Rev.* **40**, 102–113 (2011).
- Prier, C. K., Rankic, D. A. & MacMillan, D. W. C. Visible light photoredox catalysis with transition metal complexes: applications in organic synthesis. *Chem. Rev.* **113**, 5322–5363 (2013).
- Schultz, D. M. & Yoon, T. P. Solar synthesis: prospects in visible light photocatalysis. *Science* **343**, 1239176 (2014).
- Schmittel, M. & Burghart, A. Understanding reactivity patterns of radical cations. *Angew. Chem. Int. Edn Engl.* **36**, 2550–2589 (1997).
- Curran, D. P., Porter, N. A. & Giese, B. *Stereochemistry of Radical Reactions: Concepts, Guidelines, and Synthetic Applications* (VCH, 1996).
- Hopkinson, M. N., Sahoo, B., Li, J.-L. & Glorius, F. Dual catalysis sees the light: combining photoredox with organo-, acid, and transition-metal catalysis. *Chem. Eur. J.* **20**, 3874–3886 (2014).

- Nicewicz, D. A. & MacMillan, D. W. C. Merging photoredox catalysis with organocatalysis: the direct asymmetric alkylation of aldehydes. *Science* **322**, 77–80 (2008).
- Shih, H.-W., Vander Wal, M. N., Grange, R. L. & MacMillan, D. W. C. Enantioselective  $\alpha$ -benzylation of aldehydes via photoredox organocatalysis. *J. Am. Chem. Soc.* **132**, 13600–13603 (2010).
- Neumann, M., Földner, S., König, B. & Zeitler, K. Metal-free, cooperative asymmetric organophotoredox catalysis with visible light. *Angew. Chem. Int. Edn* **50**, 951–954 (2011).
- Cherevatskaya, M. *et al.* Visible-light-promoted stereoselective alkylation by combining heterogeneous photocatalysis with organocatalysis. *Angew. Chem. Int. Edn* **51**, 4062–4066 (2012).
- Nagib, D. A., Scott, M. E. & MacMillan, D. W. C. Enantioselective  $\alpha$ -trifluoromethylation of aldehydes via photoredox organocatalysis. *J. Am. Chem. Soc.* **131**, 10875–10877 (2009).
- DiRocco, D. A. & Rovis, T. Catalytic asymmetric  $\alpha$ -acylation of tertiary amines mediated by a dual catalysis mode: N-heterocyclic carbene and photoredox catalysis. *J. Am. Chem. Soc.* **134**, 8094–8097 (2012).
- Tarantino, K. T., Liu, P. & Knowles, R. R. Catalytic ketyl-olefin cyclizations enabled by proton-coupled electron transfer. *J. Am. Chem. Soc.* **135**, 10022–10025 (2013).
- Du, J., Skubi, K. L., Schultz, D. M. & Yoon, T. P. A dual-catalysis approach to enantioselective [2 + 2] photocycloadditions using visible light. *Science* **344**, 392–396 (2014).
- Bergonzini, G., Schindler, C. S., Wallentin, C.-J., Jacobsen, E. N. & Stephenson, C. R. J. Photoredox activation and anion binding catalysis in the dual catalytic enantioselective synthesis of  $\beta$ -amino esters. *Chem. Sci.* **5**, 112–116 (2013).
- Bauer, A., Westkämper, F., Grimme, S. & Bach, T. Catalytic enantioselective reactions driven by photoinduced electron transfer. *Nature* **436**, 1139–1140 (2005).
- Müller, C., Bauer, A. & Bach, T. Light-driven enantioselective organocatalysis. *Angew. Chem. Int. Edn* **48**, 6640–6642 (2009).
- Brimioulle, R. & Bach, T. Enantioselective Lewis acid catalysis of intramolecular enone [2 + 2] photocycloaddition reactions. *Science* **342**, 840–843 (2013).
- Alonso, R. & Bach, T. A chiral thioxanthone as an organocatalyst for enantioselective [2 + 2] photocycloaddition reactions induced by visible light. *Angew. Chem. Int. Edn* **53**, 4368–4371 (2014).
- Arceo, E., Jurberg, I. D., Álvarez-Fernández, A. & Melchiorre, P. Photochemical activity of a key donor–acceptor complex can drive stereoselective catalytic  $\alpha$ -alkylation of aldehydes. *Nature Chem.* **5**, 750–756 (2013).
- Huo, H., Fu, C., Harms, K. & Meggers, S. Asymmetric catalysis with substitutionally labile yet stereochemically stable chiral-at-metal iridium(III) complex. *J. Am. Chem. Soc.* **136**, 2990–2993 (2014).
- Fontecave, M., Hamelin, O. & Ménage, S. Chiral-at-metal complexes as asymmetric catalysts. *Top. Organomet. Chem.* **15**, 271–288 (2005).
- Bauer, E. B. Chiral-at-metal complexes and their catalytic applications in organic synthesis. *Chem. Soc. Rev.* **41**, 3153–3167 (2012).
- Evans, D. A., Downey, C. W. & Hubbs, J. L. Ni(II) bis(oxazoline)-catalyzed enantioselective syn aldol reactions of *N*-propionylthiazolidinethiones in the presence of silyl triflates. *J. Am. Chem. Soc.* **125**, 8706–8707 (2003).
- Sato, H. & Yamagishi, A. Application of the  $\Delta\Lambda$  isomerism of octahedral metal complexes as a chiral source in photochemistry. *J. Photochem. Photobiol. C* **8**, 67–84 (2007).
- Herrmann, A. T., Smith, L. L. & Zakarian, A. A simple method for asymmetric trifluoromethylation of *N*-acyl oxazolidinones via Ru-catalyzed radical addition to zirconium enolates. *J. Am. Chem. Soc.* **134**, 6976–6979 (2012).
- Studer, A. & Curran, D. P. The electron is a catalyst. *Nature Chem.* **6**, 765–773 (2014).
- Andrieux, C. P., Le Gorand, A. & Savéant, J. M. Electron transfer and bond breaking. Examples of passage from a sequential to a concerted mechanism in the electrochemical reductive cleavage of arylmethyl halides. *J. Am. Chem. Soc.* **114**, 6892–6904 (1992).

Supplementary Information is available in the online version of the paper.

**Acknowledgements** We acknowledge funding from the German Research Foundation (ME 1805/4-1). H.H. thanks the China Scholarship Council for a stipend.

**Author Contributions** E.M. conceived and coordinated the project and wrote the Letter. E.M. and H.H. designed the experiments. H.H. carried out the majority of the experiments. X.S. synthesized the new catalyst **A-Ir2**. C.W. contributed to the synthesis of substrates. L.Z. contributed to the synthesis and crystallization of iridium complexes. L.-A.C. provided insights into iridium enolate chemistry. P.R. performed and analysed the cyclic voltammetry under supervision of G.H. The X-ray crystallographic studies were performed by K.H. and M.M.

**Author Information** The X-ray crystallographic coordinates for structures of the iridium complex **A-Ir2**, substrate coordinated iridium complex **I** and the iridium enolate complex **II** have been deposited at the Cambridge Crystallographic Data Centre (CCDC) under deposition numbers CCDC 1014509, 1014510 and 1014876, respectively. Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to E.M. ([meggers@chemie.uni-marburg.de](mailto:meggers@chemie.uni-marburg.de)).

# Recent Northern Hemisphere stratospheric HCl increase due to atmospheric circulation changes

E. Mahieu<sup>1</sup>, M. P. Chipperfield<sup>2</sup>, J. Notholt<sup>3</sup>, T. Reddmann<sup>4</sup>, J. Anderson<sup>5</sup>, P. F. Bernath<sup>6,7,8</sup>, T. Blumenstock<sup>4</sup>, M. T. Coffey<sup>9</sup>, S. S. Dhomse<sup>2</sup>, W. Feng<sup>2</sup>, B. Franco<sup>1</sup>, L. Froidevaux<sup>10</sup>, D. W. T. Griffith<sup>11</sup>, J. W. Hannigan<sup>9</sup>, F. Hase<sup>4</sup>, R. Hossaini<sup>2</sup>, N. B. Jones<sup>11</sup>, I. Morino<sup>12</sup>, I. Murata<sup>13</sup>, H. Nakajima<sup>12</sup>, M. Palm<sup>3</sup>, C. Paton-Walsh<sup>11</sup>, J. M. Russell III<sup>5</sup>, M. Schneider<sup>4</sup>, C. Servais<sup>1</sup>, D. Smale<sup>14</sup> & K. A. Walker<sup>8,15</sup>

The abundance of chlorine in the Earth's atmosphere increased considerably during the 1970s to 1990s, following large emissions of anthropogenic long-lived chlorine-containing source gases, notably the chlorofluorocarbons. The chemical inertness of chlorofluorocarbons allows their transport and mixing throughout the troposphere on a global scale<sup>1</sup>, before they reach the stratosphere where they release chlorine atoms that cause ozone depletion<sup>2</sup>. The large ozone loss over Antarctica<sup>3</sup> was the key observation that stimulated the definition and signing in 1987 of the Montreal Protocol, an international treaty establishing a schedule to reduce the production of the major chlorine- and bromine-containing halocarbons. Owing to its implementation, the near-surface total chlorine concentration showed a maximum in 1993, followed by a decrease of half a per cent to one per cent per year<sup>4</sup>, in line with expectations. Remote-sensing data have revealed a peak in stratospheric chlorine after 1996<sup>5</sup>, then a decrease of close to one per cent per year<sup>6,7</sup>, in agreement with the surface observations of the chlorine source gases and model calculations<sup>7</sup>. Here we present ground-based and satellite data that show a recent and significant increase, at the  $2\sigma$  level, in hydrogen chloride (HCl), the main stratospheric chlorine reservoir, starting around 2007 in the lower stratosphere of the Northern Hemisphere, in contrast with the ongoing monotonic decrease of near-surface source gases. Using model simulations, we attribute this trend anomaly to a slowdown in the Northern Hemisphere atmospheric circulation, occurring over several consecutive years, transporting more aged air to the lower stratosphere, and characterized by a larger relative conversion of source gases to HCl. This short-term dynamical variability will also affect other stratospheric tracers and needs to be accounted for when studying the evolution of the stratospheric ozone layer.

Decomposition of chlorine-containing source gases in the stratosphere produces HCl, the largest reservoir of chlorine<sup>8,9</sup>. Here we investigate recent trends in atmospheric HCl with observations from eight Network for the Detection of Atmospheric Composition Change (NDACC; <http://www.ndacc.org>) ground-based stations located between 79° N and 45° S and operating Fourier Transform InfraRed (FTIR) instruments. Figure 1a shows the HCl total columns for Jungfraujoch (47° N; red squares) together with the evolution of the total tropospheric chlorine (blue curve) over the past three decades. Figure 1b–d focuses on the recent HCl changes above Ny-Ålesund (79° N) and two mid-latitude stations, Jungfraujoch (zoom of Fig. 1a) and Lauder (45° S).

At the Southern Hemisphere station we find a continuous decrease of HCl since 2001, but both Northern Hemisphere sites show an overall HCl decline, more rapid around 2004, followed by an increase from 2007 onwards. To quantify the column changes at all sites, we used a

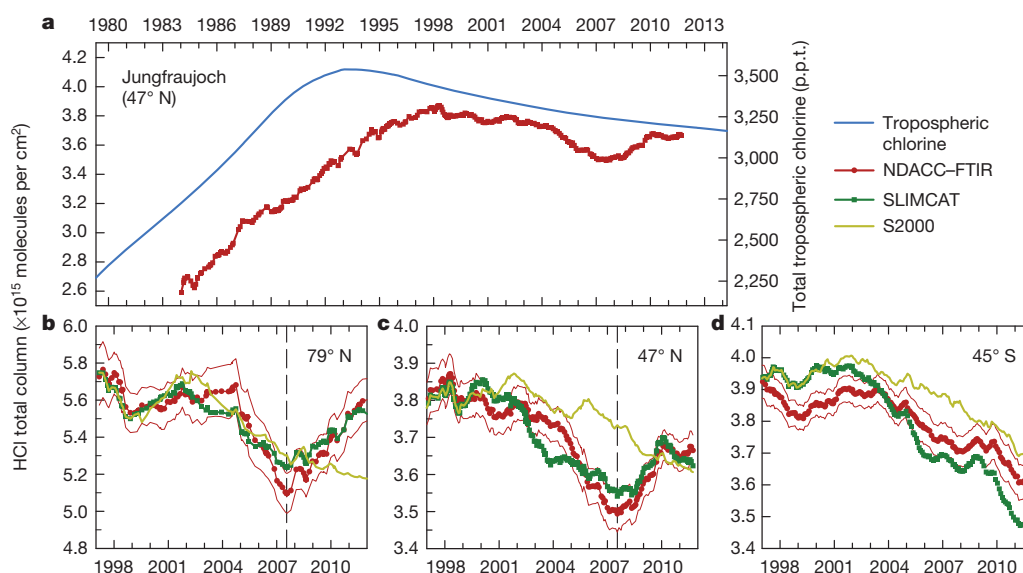
bootstrap resampling statistical tool<sup>10</sup> involving a linear component and accounting for the strong seasonal modulations present in the data sets. Figure 2 displays, for the eight NDACC sites, the relative annual HCl rates of change for the 1997–2007 and 2007–2011 time periods, using either the 1997.0 or 2007.0 computed column as reference. For the 1997–2007 time interval, we determine consistent and significant HCl decreases at all Northern Hemisphere sites, with mean relative changes ranging from –0.7 to –1.5 per cent per year. In the Southern Hemisphere, column changes are not significant at the  $2\sigma$  level. For 2007–2011, mean relative column growths of 1.1–3.4 per cent per year are derived for all Northern Hemisphere sites while negative or undefined rates are observed for Wollongong and Lauder in the Southern Hemisphere.

To corroborate these findings with independent data, and to get information on the altitude range where these changes occur, we included the GOZCARDS<sup>11</sup> satellite data set (Global OZone Chemistry And Related Data sets for the Stratosphere version 1.01), which merges observations by the HALOE<sup>12</sup> (HALogen Occultation Experiment version 19), ACE-FTS<sup>13</sup> (Atmospheric Chemistry Experiment-Fourier Transform Spectrometer version 2.2) and Aura/MLS<sup>14</sup> (Microwave Limb Sounder version 3.3) instruments. Partial columns were computed between 100 hPa and 10 hPa, considering the zonal monthly mean mixing ratio time series available for the whole time interval in the 70°–80° N, 60°–70° N, 40°–50° N, 30°–40° N, 20°–30° N, 30°–40° S and 40°–50° S latitudinal bands. These partial columns typically span altitudes of 16–31 km, that is, the region with maximum HCl concentration and in which the FTIR measurements are most sensitive<sup>5</sup>.

Corresponding rates of change are also displayed in Fig. 2. For 1997–2007, there is excellent agreement in the Northern Hemisphere between the satellite and the six NDACC-FTIR trends determined above. In the Southern Hemisphere, GOZCARDS reveals statistically significant decreases of HCl at the  $2\sigma$  level, while the FTIR time series suggest stable columns at the same level of confidence. For 2007–2011, the ACE-FTS and Aura/MLS merged data confirm the upward FTIR trends in the Northern Hemisphere. Figure 3 illustrates this, showing satellite monthly means (red dots) for 30°–60° N and 30°–60° S, at 46 hPa and 7 hPa, together with a linear fit to the data for both time periods. The HCl increase is clearly confined to the Northern Hemisphere lower stratosphere.

Because HCl is the main final product of the decomposition of any chlorine-containing source gases, we need to verify that its rise after 2007 does not result from the substantial contribution of new unknown sources of chlorine whose emissions occur predominantly in the Northern Hemisphere, not monitored by the *in situ* networks, and unregulated by the Montreal Protocol, its Amendments and Adjustments. Indeed, such chlorine-containing source gases have been recently identified<sup>15</sup>,

<sup>1</sup>Institute of Astrophysics and Geophysics, University of Liège, Liège 4000, Belgium. <sup>2</sup>National Centre for Atmospheric Science, School of Earth and Environment, University of Leeds, Leeds LS2 9JT, UK. <sup>3</sup>Department of Physics, University of Bremen, Bremen 28334, Germany. <sup>4</sup>Karlsruhe Institute of Technology (KIT), Institute for Meteorology and Climate Research (IMK-ASF), Karlsruhe 76021, Germany. <sup>5</sup>Department of Atmospheric and Planetary Science, Hampton University, Hampton, Virginia 23668, USA. <sup>6</sup>Department of Chemistry and Biochemistry, Old Dominion University, Norfolk, Virginia 23529, USA. <sup>7</sup>Department of Chemistry, University of York, York YO10 5DD, UK. <sup>8</sup>Department of Chemistry, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada. <sup>9</sup>National Center for Atmospheric Research, Boulder, Colorado 80307, USA. <sup>10</sup>Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California 91109, USA. <sup>11</sup>School of Chemistry, University of Wollongong, Wollongong, New South Wales 2522, Australia. <sup>12</sup>National Institute for Environmental Studies (NIES), Tsukuba, Ibaraki 305-8506, Japan. <sup>13</sup>Graduate School of Environmental Studies, Tohoku University, Sendai 980-8578, Japan. <sup>14</sup>National Institute of Water and Atmospheric Research (NIWA), Lauder 9352, New Zealand. <sup>15</sup>Department of Physics, University of Toronto, Toronto, Ontario M5S 1A7, Canada.



**Figure 1 | Evolution of HCl in the Earth's atmosphere.** **a**, The long-term total column time series of HCl at Jungfraujoch (running average with a 3-yr integration length, step of 1 month; in red, left scale) and the global total tropospheric chlorine volume mixing ratio (blue curve, right scale, in parts per trillion, p.p.t.). The lower panels display the running average total column time

series (1997–2011) of HCl at Ny-Ålesund (**b**), Jungfraujoch (**c**) and Lauder (**d**), derived from the NDACC-FTIR observations, and the standard (green) and S2000 (light green) SLIMCAT simulations. The thin red lines correspond to the  $\pm 2$  standard error of the mean range. Minimum columns are observed in July 2007 at the Northern Hemisphere sites (dashed lines).

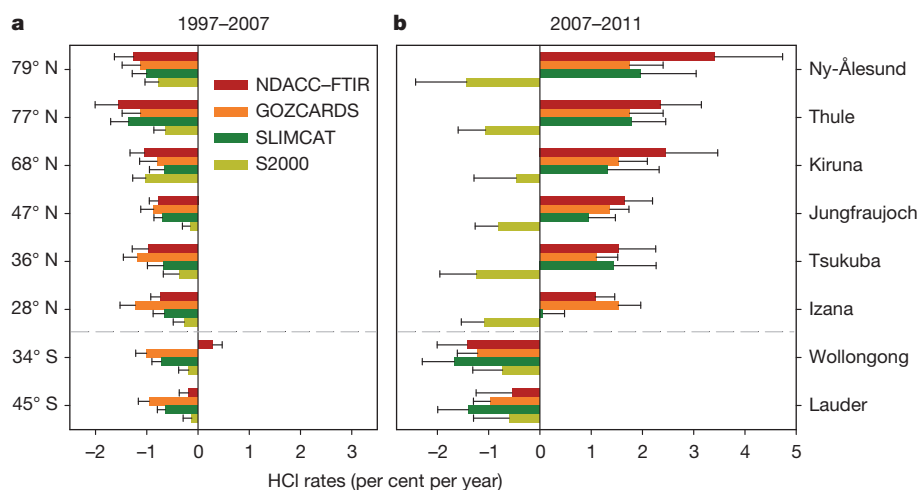
although in that case, their contribution to the HCl upturn can be ruled out by their very low concentrations.

We have used results from two state-of-the-art three-dimensional chemical transport models, SLIMCAT<sup>7</sup> and KASIMA<sup>7</sup>, to interpret the recent HCl increase. Both models performed a standard simulation using surface source gas mixing ratios from the WMO A1 (World Meteorological Organisation; 2010) emission scenario<sup>4</sup> and were forced using ERA-Interim meteorological fields<sup>16</sup> from the European Centre for Medium-Range Weather Forecasts (ECMWF). The key results for HCl trends from both models agree. Here we show data from the SLIMCAT runs; corresponding results from KASIMA are shown in Extended Data Figs 1–4. To study the impact of atmospheric dynamics, an additional SLIMCAT run (S2000) used constant 2000 meteorological forcing, from 2000 onwards.

Running averages for both SLIMCAT simulations are reproduced in Fig. 1b–d. For the three sites, run S2000 (light green curve) predicts an overall HCl decrease while the standard run (green squares) reproduces the observed and distinct evolution prevailing in both hemispheres, after correction of a constant low-bias of about 7% in the Northern Hemisphere simulations. The total column changes characterizing the model data

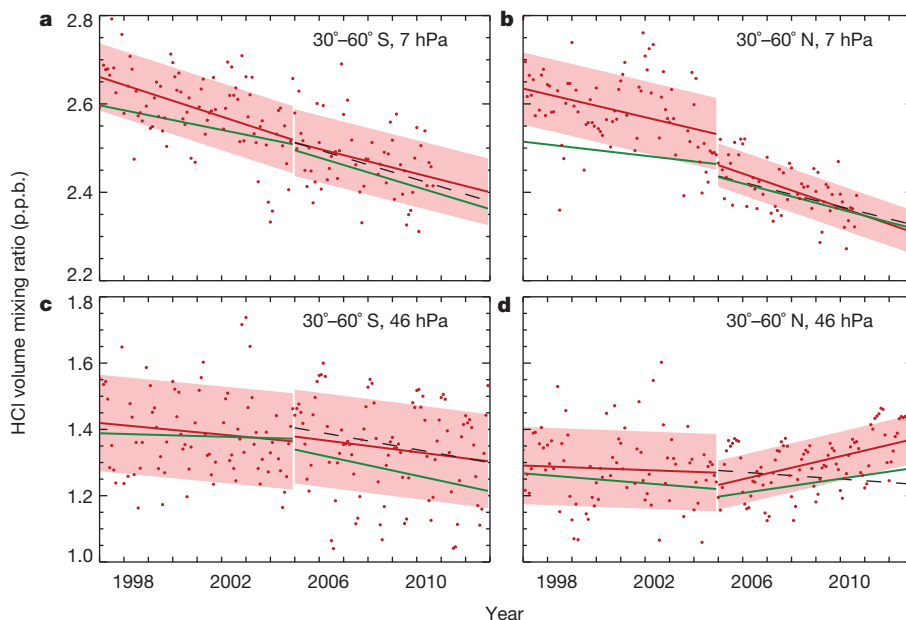
sets are displayed in Fig. 2. The model runs predict significant decreases in HCl for the 1997–2007 reference period at all sites and there is an overall agreement within the error bars for the amplitude of the signals between the model and the observations. Regarding the 2007–2011 time period, the SLIMCAT time series are characterized by positive trends from Ny-Ålesund (79° N) to Tsukuba (36° N) and by significant decreases for the Southern Hemisphere stations, but show no change for the near-tropical site of Izana (28° N). The S2000 sensitivity run does not produce the HCl trend reversal and, instead, indicates declines at all sites.

The agreement between measurement and model demonstrates that the HCl increase after 2007 is not caused by new, unidentified chlorine sources, or by underestimates in emissions of known species of chlorine-containing source gases, because these are used as model input. The agreement between model and observation also shows that there is a good understanding of the chemistry which converts source gases to HCl. The difference between the HCl trends forecast by the two SLIMCAT runs—that is, a significant increase for northern high- and mid-latitudes or a constant decrease below 30° N—establishes that changes in the atmospheric circulation cause the recent HCl increase, since only the



**Figure 2 | HCl relative rates of change for eight NDACC sites.** **a**, The rates of change (per cent per year) for the 1997–2007 time period (1999–2007 for Thule and Izana, 1998–2007 for Tsukuba). **b**, As for **a** but for 2007–2011. The rates of change were derived from the FTIR and GOZCARDS observational data sets and from the two SLIMCAT simulated time series (see colour key). The error bars correspond to the  $2\sigma$  level of uncertainty.





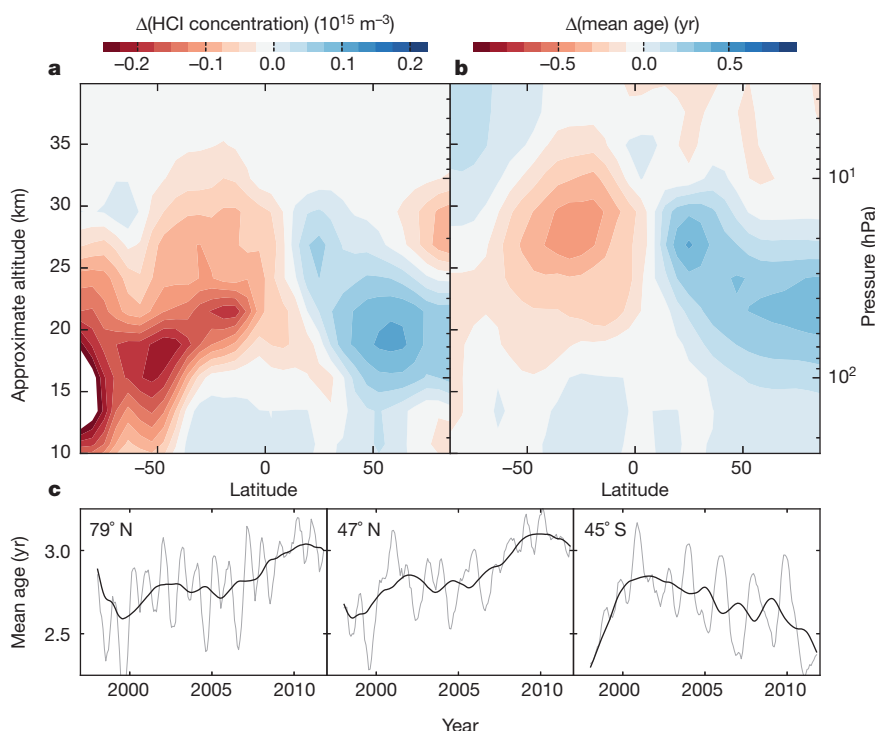
**Figure 3 | Evolution of stratospheric HCl from satellite observations.** Comparison of merged GOZCARDs satellite HCl observations (by HALOE, ACE-FTS and Aura/MLS) with SLIMCAT model runs for Northern Hemisphere and Southern Hemisphere mid-latitude lower (46 hPa) and upper (7 hPa) stratosphere. GOZCARDs monthly means are shown as red dots. Linear fits to the GOZCARDs data and standard SLIMCAT run are displayed as red and green lines, respectively, for periods before and after 2005. The dashed black line shows fits to the S2000 run, which assumes no change in circulation. An upward trend is observed in the Northern Hemisphere lower stratosphere (d) while HCl is decreasing in the southern and northern upper stratosphere (a, b); volume mixing ratio in parts per billion (p.p.b.).

meteorological fields adopted from 2000 onwards differ between the two runs. To diagnose these circulation changes, we examined age-of-air maps produced by the standard SLIMCAT run. They reveal a slower circulation in the Northern Hemisphere lower stratosphere after 2005–2006, with older air characterized by a larger relative conversion of the chlorine-containing source gases into HCl.

Figure 4b shows the age-of-air change between 2005–2006 and 2010–2011. Air older by up to 0.4 yr is found at altitudes of around 20–25 km in a broad range of Northern Hemisphere latitudes, in a region where the mean age-of-air is typically about 3 yr. There is an obvious correlation with the evolution of the HCl concentrations over the same time period (Fig. 4a), which exhibits a very similar pattern and hemispheric asymmetry. Time series of mean age-of-air near 50 hPa above Ny-Ålesund, Jungfraujoch and Lauder are displayed in Fig. 4c. The 3-yr running

means (black curves) indicate a progressive slowdown of the Northern Hemisphere stratospheric circulation after 2005–2006. For Lauder, a fairly constant circulation speedup occurs from 2000 onwards.

These changes are significant at the  $2\sigma$  level, with Northern Hemisphere air ageing by 3–4 weeks per year after 2005, compared to about 1 week per year before. For Lauder, the mean age-of-air change during the last decade is calculated to be  $-2$  weeks per year. Other important factors, such as the details of specific transport pathways, which lead to a given mean age-of-air, also affect the conversion rate of the source gases to HCl (ref. 17). These pathways are simulated by the model but not revealed by the simple diagnostic of mean age-of-air. The slower Northern Hemisphere circulation occurring over a few years after 2005–2006 seems to contrast with the speedup of the Brewer–Dobson circulation, which is predicted in the very long-term to be a response to climate



**Figure 4 | Spatial distribution of the HCl concentration and age-of-air changes.** Mean differences of the HCl concentration (a) and age-of-air (b) between 2010/2011 and 2005/2006, as a function of altitude and latitude, derived from the standard SLIMCAT simulation. There is a clear asymmetry between the hemispheres, with correlated patterns between age-of-air and HCl, indicating that the HCl changes over that period are consistent with slower/faster circulation in the Northern/Southern Hemisphere. c, Running averages of the mean age-of-air at 50 hPa (thick/thin curve, integration length of 36/6 months), at the same sites as Fig. 1 (the time series at 79° N and 45° S have been shifted vertically by  $-0.75$  yr).

change<sup>18,19</sup>, but the recent slowdown is probably part of dynamical variability occurring on shorter timescales: it does not imply a change in the general circulation strength. More than year-to-year variability, it is multiyear periods of age-of-air increase or decrease, such as those highlighted in our study or reported recently<sup>20</sup>, that will probably complicate the search of a long-term trend in mean circulation.

We have presented observations and simulations of a recent HCl increase in the Northern Hemisphere lower stratosphere. We ascribe it to dynamical variability, occurring on a timescale of a few years, characterized by a persistent slowing of stratospheric circulation after 2005, bringing HCl-enriched air into the Northern Hemisphere lower stratosphere. We find no evidence that unidentified chlorine-containing source gases are responsible for this HCl increase. In the Southern Hemisphere, a fairly constant decrease has been observed over the past ten years. Globally, our ground-based observations indicate a mean HCl decrease of 0.5 per cent per year for 1997–2011, compatible with the 0.5–1 per cent per year range that characterized the post-peak reduction of tropospheric chlorine<sup>4</sup>. Hence, we conclude that the Montreal Protocol is still on track, and is leading to an overall reduction of the stratospheric chlorine loading. However, multiyear variability in the stratospheric circulation and dynamics, as identified here, could lead to further unpredictable increases or redistribution of HCl and other stratospheric tracers. Therefore, such variability and its causes will have to be thoroughly characterized and carefully accounted for when evaluating trends or searching for ozone recovery.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 27 March; accepted 10 September 2014.**

- Lovelock, J. E., Maggs, R. J. & Wade, R. J. Halogenated hydrocarbons in and over the Atlantic. *Nature* **241**, 194–196 (1973).
- Molina, M. J. & Rowland, F. S. Stratospheric sink for chlorofluoromethanes: chlorine atom-catalysed destruction of ozone. *Nature* **249**, 810–812 (1974).
- Farman, J. C., Gardiner, B. G. & Shanklin, J. D. Large losses of total ozone in Antarctica reveal seasonal ClO<sub>x</sub>/NO<sub>x</sub> interaction. *Nature* **315**, 207–210 (1985).
- World Meteorological Organization *Scientific Assessment of Ozone Depletion: 2010* (Report 52, Global Ozone Research and Monitoring Project, WMO, 2011); [http://www.wmo.int/pages/prog/arep/gaw/ozone\\_2010/ozone\\_asst\\_report.html](http://www.wmo.int/pages/prog/arep/gaw/ozone_2010/ozone_asst_report.html).
- Rinsland, C. P. *et al.* Long-term trends of inorganic chlorine from ground-based infrared solar spectra: past increases and evidence for stabilization. *J. Geophys. Res.* **108** (D8), 27, <http://dx.doi.org/10.1029/2002JD003001> (2003).
- Froidevaux, L. *et al.* Temporal decrease in upper atmospheric chlorine. *Geophys. Res. Lett.* **33**, <http://dx.doi.org/10.1029/2006GL027600> (2006).
- Kohlhepp, R. *et al.* Observed and simulated time evolution of HCl, ClONO<sub>2</sub>, and HF total column abundances. *Atmos. Chem. Phys.* **12**, 3527–3556 (2012).
- Zander, R. *et al.* The 1985 chlorine and fluorine inventories in the stratosphere based on ATMOS observations at 30° north latitudes. *J. Atmos. Chem.* **15**, 171–186 (1992).
- Nassar, R. *et al.* A global inventory of stratospheric chlorine in 2004. *J. Geophys. Res.* **111**, D22312, <http://dx.doi.org/10.1029/2006JD007073> (2006).
- Gardiner, T. *et al.* Trend analysis of greenhouse gases over Europe measured by a network of ground-based remote FTIR instruments. *Atmos. Chem. Phys.* **8**, 6719–6727 (2008).
- Froidevaux, L. *et al.* GOZCARDS Merged Data for Hydrogen Chloride Monthly Zonal Means on a Geodetic Latitude and Pressure Grid version 1.01, <http://dx.doi.org/10.5067/MEASURES/GOZCARDS/DATA3002> (NASA Goddard Earth Science Data and Information Services Center, accessed June 2013).
- Russell, J. M. III *et al.* The halogen occultation experiment. *J. Geophys. Res.* **98**, 10777–10797 (1993).
- Bernath, P. F. *et al.* Atmospheric Chemistry Experiment (ACE): mission overview. *Geophys. Res. Lett.* **32**, L15S01, <http://dx.doi.org/10.1029/2005GL022386> (2005).
- Waters, J. W. *et al.* The Earth Observing System Microwave Limb Sounder (EOS MLS) on the Aura satellite. *IEEE Trans. Geosci. Rem. Sens.* **44**, 1075–1092 (2006).
- Laube, J. C. *et al.* Newly detected ozone-depleting substances in the atmosphere. *Nature Geosci.* **7**, 266–269 (2014).
- Dee, D. P. *et al.* The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Q. J. R. Meteorol. Soc.* **137**, 553–597 (2011).
- Wagh, D. W., Strahan, S. E. & Newman, P. A. Sensitivity of stratospheric inorganic chlorine to differences in transport. *Atmos. Chem. Phys.* **7**, 4935–4941 (2007).
- Engel, A. *et al.* Age of stratospheric air unchanged within uncertainties over the past 30 years. *Nature Geosci.* **2**, 28–31 (2009).
- McLandress, C. & Shepherd, T. G. Simulated anthropogenic changes in the Brewer-Dobson circulation, including its extension to high latitude. *J. Clim.* **22**, 1516–1540 (2009).
- Stiller, G. P. *et al.* Observed temporal evolution of global mean age of stratospheric air for the 2002 to 2010 period. *Atmos. Chem. Phys.* **12**, 3311–3331 (2012).
- Rothman, L. S. *et al.* The HITRAN 2008 molecular spectroscopic database. *J. Quant. Spec. Radiat. Transf.* **110**, 533–572 (2009).

**Acknowledgements** The University of Liège contribution was mainly supported by the Belgian Science Policy Office (BELSPO) and the Fonds de la Recherche Scientifique-FNRS, both in Brussels. Additional support was provided by MeteoSwiss (Global Atmospheric Watch) and the Fédération Wallonie-Bruxelles. We thank the International Foundation High Altitude Research Stations Jungfraujoch and Gornergrat (HFSJG, Bern). We thank O. Flock and D. Zander (University of Liège). The SLIMCAT modelling work was supported by the UK Natural Environment Research Council (NCAS and NCEO). The FTIR measurements at Ny-Ålesund, Spitsbergen, are supported by the AWI Bremerhaven. The work from Hampton University was partially funded under the NASA MEASURE's GOZCARDS programme and the National Oceanic and Atmospheric Administration's Educational Partnership Program Cooperative Remote Sensing Science and Technology Center (NOAA EPP CREST). The ACE mission is supported primarily by the Canadian Space Agency. We thank U. Raffalski and P. Voelger for technical support at IRF Kiruna. The National Center for Atmospheric Research is supported by the National Science Foundation. The observation programme at Thule, Greenland, is supported under contract by the National Aeronautics and Space Administration (NASA) and the site is also supported by the NSF Office of Polar Programs. We thank the Danish Meteorological Institute for support at Thule. Work at the Jet Propulsion Laboratory, California Institute of Technology, was performed under contract with NASA; we thank R. Fuller for help in producing the GOZCARDS data set, and work by many ACE-FTS, HALOE and MLS team members who helped to produce data towards the GOZCARDS data set is also acknowledged. We thank O. E. García, E. Sepúlveda, and the State Meteorological Agency (AEMET) of Spain for scientific and technical support at Izana. The Australian Research Council has provided notable support over the years for the NDACC site at Wollongong, most recently as part of project DP110101948. Measurements at Lauder are core funded through New Zealand's Ministry of Business, Innovation and Employment. We are grateful to all colleagues who have contributed to FTIR data acquisition. We thank ECMWF for providing the ERA-Interim reanalyses.

**Author Contributions** M.P., J.W.H., F.H., E.M., I. Murata, N.B.J., C.P.-W. and D.S. performed the Ny-Ålesund, Thule, Kiruna and Izana, Jungfraujoch, Tsukuba, Wollongong and Lauder retrievals for HCl, respectively. P.F.B. and K.A.W. provided ACE-FTS data; L.F. and J.A. provided the GOZCARDS data set. J.A., P.F.B., L.F., J.M.R. III and K.A.W. provided expertise on satellite data usage. M.P.C., R.H., S.S.D. and W.F. designed and performed the SLIMCAT runs, sensitivity analyses and transport diagnostics. T.R. performed the KASIMA model run and corresponding diagnostics. B.F. and E.M. performed the trend analyses and compiled the results. J.N., M.T.C., T.B., C.S., I. Morino, H.N., M.S., D.W.T.G. and D.S. are responsible for the instrumentation and data acquisition at the NDACC stations. E.M. initiated and coordinated the study. The figures were prepared by E.M. and B.F. (Fig. 1), E.M. (Fig. 2), R.H. and M.P.C. (Fig. 3) and T.R. (Fig. 4). E.M., M.P.C. and J.N. wrote the manuscript. Together with T.R., they revised it and included the comments from the co-authors.

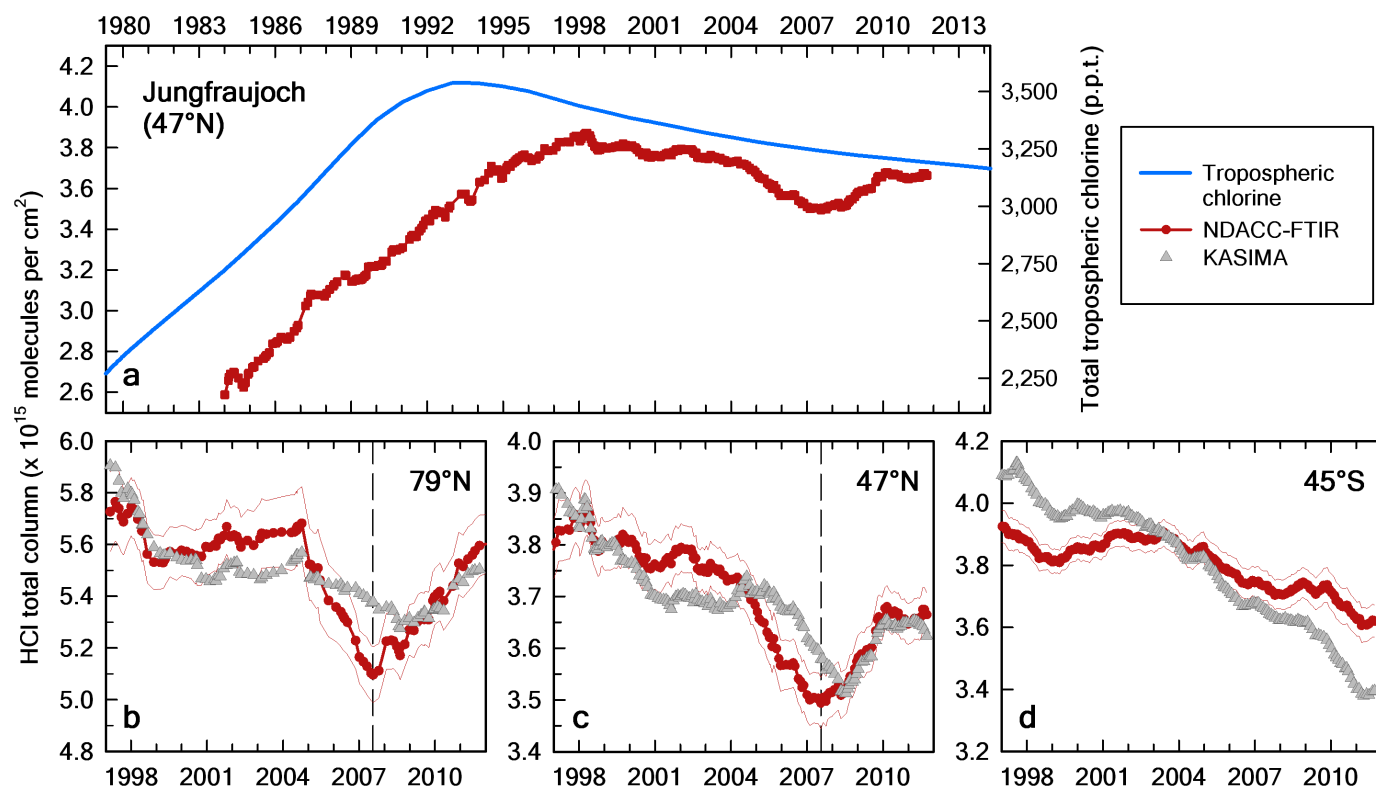
**Author Information** NDACC data are publicly available at <ftp://ftp.cpc.ncep.noaa.gov/ndacc/station/> and GOZCARDS data are publicly available at <http://measures.gsfc.nasa.gov/opendap/GOZCARDS/>. Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to E.M. (emmanuel.mahieu@ulg.ac.be).

## METHODS

The ground-based observations were performed at the NDACC sites by solar absorption spectrometry in the infrared spectral region, using FTIR high-resolution instruments. Observations are recorded under clear sky conditions year-round, except at Ny-Ålesund and Thule, where the polar night prevents measurements between about October and February. The HCl total columns were retrieved with the SFIT-2, SFIT-4 or PROFFIT algorithm in narrow spectral ranges encompassing isolated lines of HCl<sup>5,7</sup>, generally assuming pressure-temperature profiles provided by the National Centers for Environmental Prediction (NCEP). The GOZCARDS<sup>11</sup> data set for HCl includes zonal average monthly mean time series of stratospheric mixing ratio profiles merging individual measurements from the HALOE (1991–2005), ACE-FTS (2004 onward) and Aura MLS (2004 onward) satellite-borne instruments. Line parameters from recent HITRAN databases<sup>21</sup> were adopted in the spectrometric

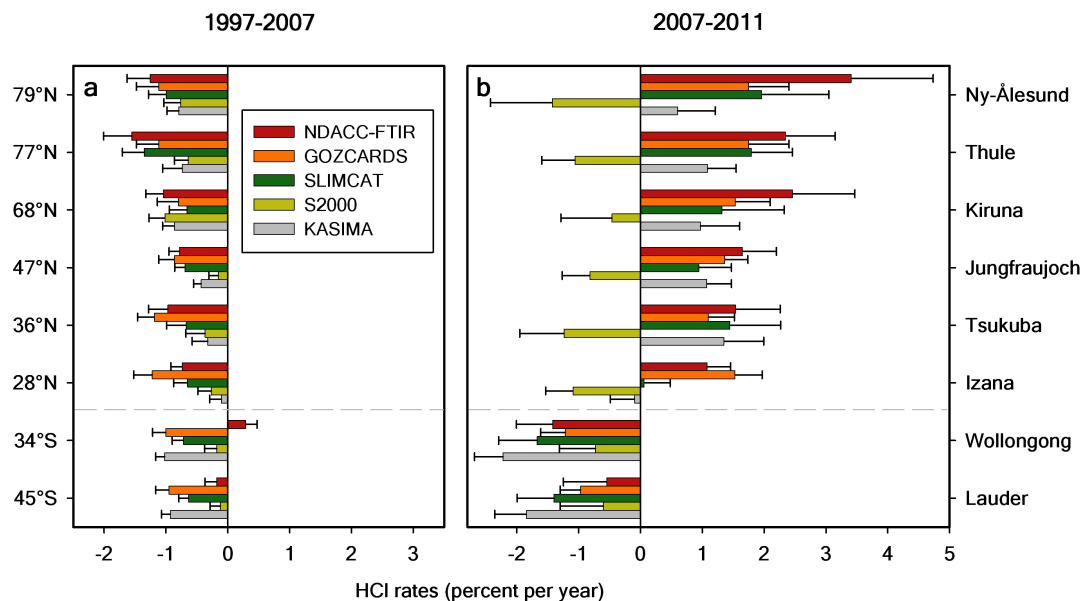
analyses. We used the SLIMCAT and KASIMA models<sup>7</sup> to support our investigations. Both used ERA-Interim analyses provided by ECMWF<sup>16</sup>, and they provided consistent results for the HCl trends, giving confidence in their robustness. The models contain detailed treatments of stratospheric chemistry and have been extensively used for studies of stratospheric ozone<sup>7</sup>. Stratospheric age-of-air was diagnosed in the model runs using an idealized tracer with a linearly increasing tropospheric mixing ratio. For the S2000 SLIMCAT simulation, 6-hourly winds of 2000 were used every year from 2000 onwards. The trend determinations were performed with a bootstrap resampling statistical tool<sup>10</sup>, considering all available daily or monthly means (excluding the winter months for the very high-latitude sites) while the model data sets were limited to days with available FTIR measurements. We studied the impact of the FTIR sampling using the bootstrap algorithm, and found no statistically significant impact on the calculated trends.





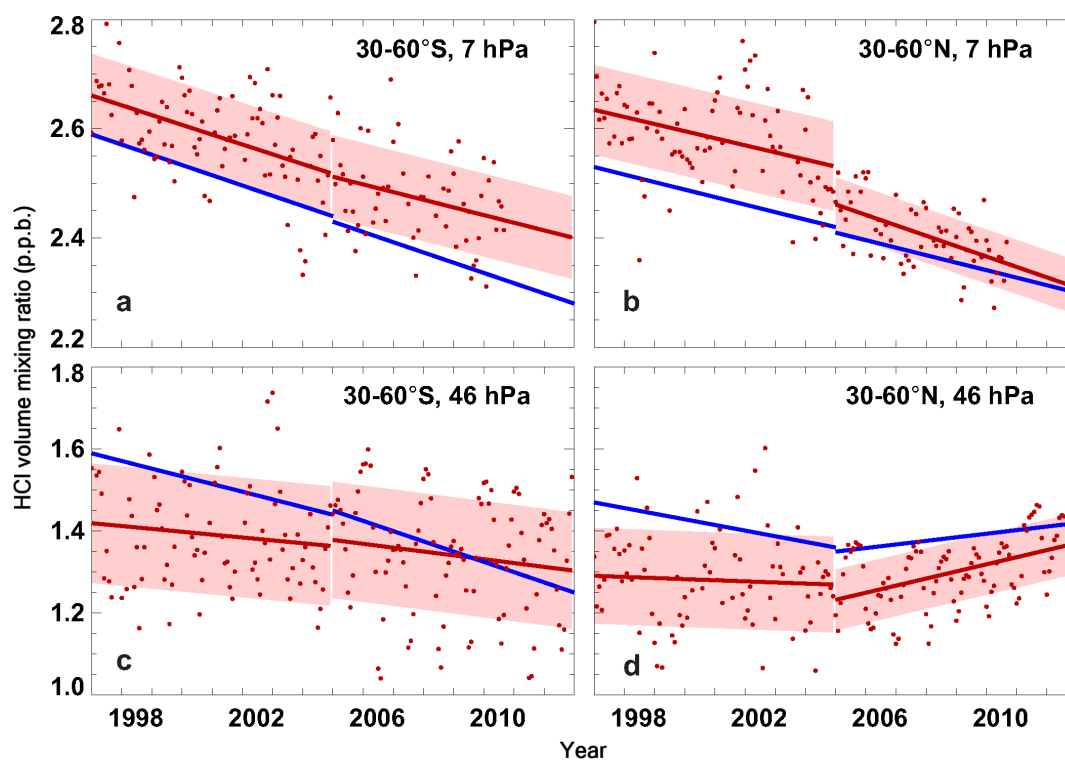
**Extended Data Figure 1 | Evolution of HCl in the Earth's atmosphere and comparison with KASIMA model results.** **a**, The long-term total column time series of HCl at Jungfraujoch (running average with a 3-yr integration length, step of 1 month; in red, left scale, in molecules per  $\text{cm}^2$ ) and the global total tropospheric chlorine volume mixing ratio (blue curve, right scale). Lower panels display the running average total column time series (1997–2011) of HCl

at Ny-Ålesund (**b**), Jungfraujoch (**c**) and Lauder (**d**), derived from the NDACC-FTIR observations and from the KASIMA run (grey). The thin red lines correspond to the  $\pm 2$  standard error of the mean range. The vertical dashed lines identify the occurrence of the minimum total columns at the Northern Hemisphere sites, in July 2007.



**Extended Data Figure 2 | HCl relative rates of change at eight NDACC sites.** **a** and **b** provide the rates of change (per cent per year) for the 1997–2007 (1999–2007 for Thule and Izana, 1998–2007 for Tsukuba) and 2007–2011 time

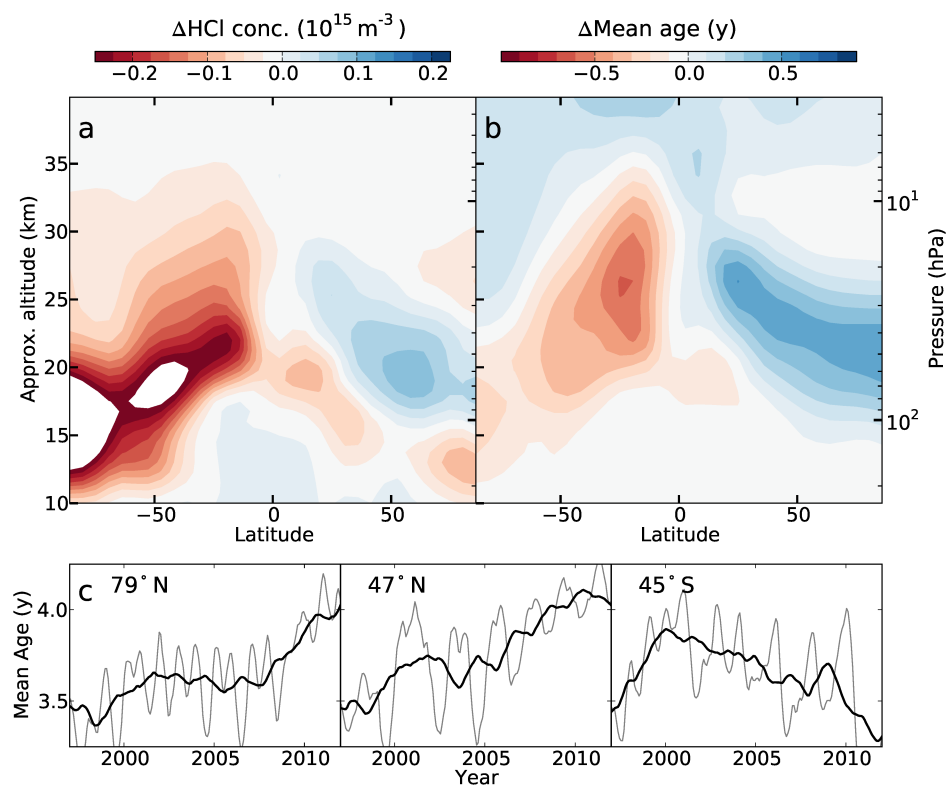
periods, respectively. They were derived from the FTIR and GOZCARDS observational data sets and from the SLIMCAT and KASIMA simulated time series (see colour key). The error bars correspond to the  $2\sigma$  level of uncertainty.



**Extended Data Figure 3 | Evolution of stratospheric HCl from satellite observations.** Comparison of merged GOZCARDS satellite HCl observations (by HALOE, ACE-FTS and Aura/MLS) with KASIMA model results for Northern and Southern Hemisphere mid-latitude lower (46 hPa) and upper (7 hPa) stratosphere. GOZCARDS monthly mean observations are shown as

red dots. Linear fits to the GOZCARDS data and the KASIMA run are displayed as red and blue lines, respectively, for periods before and after 2005. An upward trend is observed and modelled in the Northern Hemisphere lower stratosphere (**d**) while HCl is decreasing in the southern and northern upper stratosphere (**a**, **b**); volume mixing ratio in parts per billion.





**Extended Data Figure 4 | Spatial distribution of the HCl concentration and age-of-air changes.** Mean differences of the HCl concentration (a) and age-of-air (b) between 2010/11 and 2005/06, as a function of altitude and latitude, derived from the KASIMA model simulation. c, Running averages of the mean age-of-air at 50 hPa (thick/thin curve, integration length of 36/6 months), at the

same sites as in Fig. 1 (time series at 79° N/45° S have been shifted vertically by –0.75/–0.50 yr). Comparison with age-of-air time series derived from SLIMCAT (see Fig. 4c) indicates that KASIMA provides higher absolute values of mean age-of-air. Note that the upper boundary of KASIMA is at 120 km, yielding higher mean ages, compared to SLIMCAT (upper boundary 60 km).

# Selection for niche differentiation in plant communities increases biodiversity effects

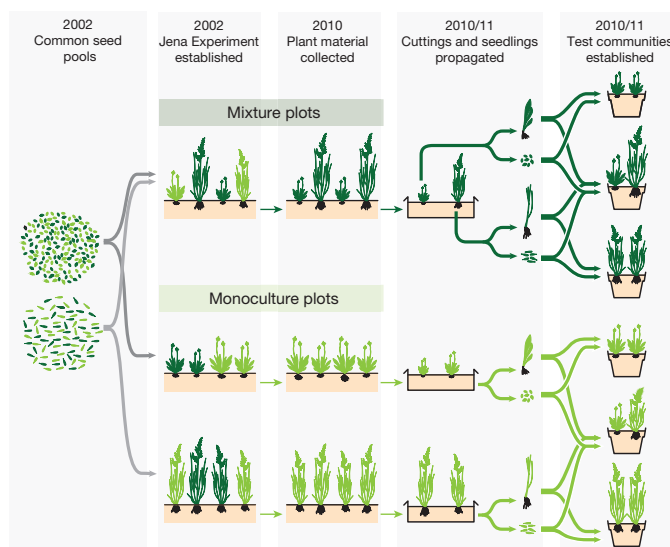
Debra Zuppinge-Dingley<sup>1</sup>, Bernhard Schmid<sup>1</sup>, Jana S. Petermann<sup>2,3</sup>, Varuna Yadav<sup>1</sup>, Gerlinde B. De Deyn<sup>4</sup> & Dan F. B. Flynn<sup>1,5</sup>

In experimental plant communities, relationships between biodiversity and ecosystem functioning have been found to strengthen over time<sup>1,2</sup>, a fact often attributed to increased resource complementarity between species in mixtures<sup>3</sup> and negative plant–soil feedbacks in monocultures<sup>4</sup>. Here we show that selection for niche differentiation between species can drive this increasing biodiversity effect. Growing 12 grassland species in test monocultures and mixtures, we found character displacement between species and increased biodiversity effects when plants had been selected over 8 years in species mixtures rather than in monocultures. When grown in mixtures, relative differences in height and specific leaf area between plant species selected in mixtures (mixture types) were greater than between species selected in monocultures (monoculture types). Furthermore, net biodiversity and complementarity effects<sup>1,2</sup> were greater in mixtures of mixture types than in mixtures of monoculture types. Our study demonstrates a novel mechanism for the increase in biodiversity effects: selection for increased niche differentiation through character displacement. Selection in diverse mixtures may therefore increase species coexistence and ecosystem functioning in natural communities and may also allow increased mixture yields in agriculture or forestry. However, loss of biodiversity and prolonged selection of crops in monoculture may compromise this potential for selection in the longer term.

Higher biodiversity promotes stability and productivity, with an increasing effect over time<sup>1,2</sup>. These positive biodiversity effects on stability and productivity can arise from complementarity between species in resource use, such as partitioning of soil resources<sup>5,6</sup> or, in some cases, accumulation of greater resources at high diversity<sup>7</sup>. Concurrently, the accumulation of natural enemies in low diversity, the Janzen–Connell effect<sup>8,9</sup>, both promotes species coexistence through density-dependent mortality and limits productivity owing to the high pressure of species-specific pathogens<sup>4</sup>. Increasing biodiversity effects could therefore arise from increasingly complementary resource use and nutrient accumulation in mixtures or pathogen accumulation over time in monocultures. Here we propose a distinct, novel mechanism—that increased biodiversity effects over time result from selection for increased niche differentiation<sup>5,6</sup> between plant species in diverse plant communities, reducing competition between species. Character displacement<sup>10</sup>, as reflected in functional trait differences, may drive such increasing niche differentiation between species<sup>11</sup>. Our hypothesis predicts larger functional trait differences between species in mixtures and, associated with such divergence in traits, stronger positive biodiversity effects on productivity. Selection for increased niche differentiation could thus explain the experimentally observed<sup>1,2</sup> increasing biodiversity effects over time, and may have implications for the effect of biodiversity on agricultural and forestry production, in which genetic diversity is known to promote production<sup>12</sup> and biodiversity may be essential in maintaining the pace of production gains<sup>13</sup>.

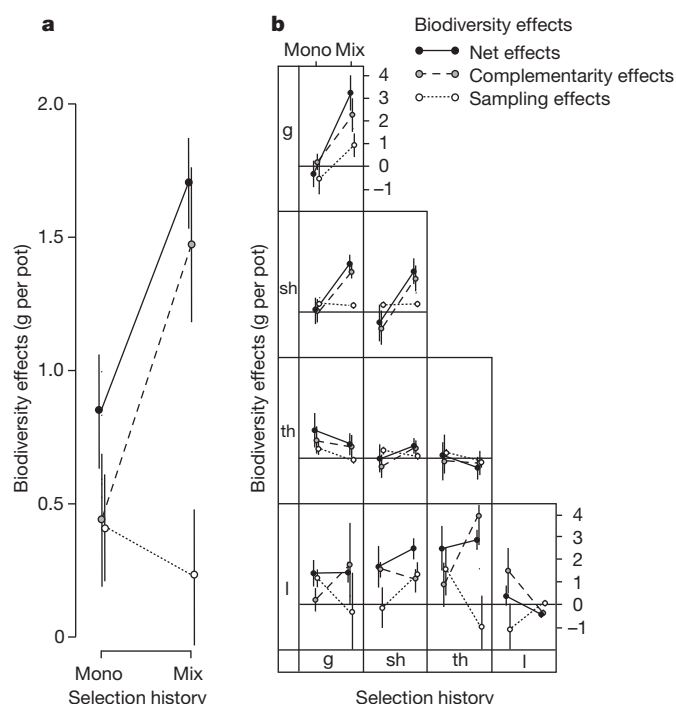
According to our hypothesis, mixture communities composed of the progeny of plants grown in conditions of high diversity ('mixture types') should have greater complementarity than mixture communities

composed of the progeny of plants grown in monocultures ('monoculture types'). We hypothesize that heritable phenotypes resulting from distinct genotypic or epigenetic features, including maternal effects, can arise from selection pressures in diverse communities. Our hypothesis of selection for niche differentiation through character displacement predicts adaptation to local diversity, with mixture types selected for niche differentiation and high performance in mixtures. We tested our hypothesis by growing mixture and monoculture types of 12 plant species—collected from selection communities of monocultures, mixtures of four or more species of a single functional group, and mixtures of four or more species of four functional groups (grasses, small herbs, tall herbs and legumes)—in pots containing four individuals of one or two species in a glasshouse (Extended Data Fig. 1). Our selection communities (selection history) were experimental plots of an 8-year biodiversity field experiment in Jena, Germany<sup>14</sup>, and our test communities (planted diversity) were the pots containing four individuals each (Fig. 1). We assessed above-ground plant biomass in mixtures versus monocultures to calculate biodiversity effects, partitioned into complementarity and sampling effects<sup>15</sup> (which are usually called selection effects<sup>15</sup>, a phrase we will not use here to avoid confusion). We consider higher net biodiversity and complementarity effects for mixture types as an indication of niche differentiation,



**Figure 1 | Experimental design.** Plant material, shoots and roots ( $n = 4,900$ ) from experimental field monoculture versus mixture selection communities established in Jena, Germany in 2002, were collected in 2010 and assembled in new experimental glasshouse monoculture versus mixture test communities in 2010/11 ( $n = 855$ ). We expected that mixture test communities would have higher productivity if assembled from plants collected from mixture selection communities in the field and vice versa for monocultures. Different shades of green represent the hypothesized selection for monoculture or mixture types from 2002 to 2010.

<sup>1</sup>Institute of Evolutionary Biology and Environmental Studies & Zurich-Basel Plant Science Center, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland. <sup>2</sup>Institute of Biology, Freie Universität Berlin, Königin-Luise-Str. 1-3, 14195 Berlin, Germany. <sup>3</sup>Berlin-Brandenburg Institute of Advanced Biodiversity Research (BBIB), 14195 Berlin, Germany. <sup>4</sup>Environmental Sciences, University of Wageningen, Droevendaalsesteeg 4, 6708PB Wageningen, the Netherlands. <sup>5</sup>Arnold Arboretum, Harvard University, Boston, Massachusetts 02131, USA.



**Figure 2 | Stronger biodiversity effects for plants selected in mixtures compared with plants selected in monocultures.** **a**, **b**, Plants selected in mixture plots in the Jena Experiment over 8 years showed stronger biodiversity effects than plants selected in monoculture plots over the same time period. Mix, mixture types; Mono, monoculture types. **a**, Biodiversity effects were assessed by additive partitioning<sup>15</sup> of net effects into complementarity and sampling effects ( $n = 545$ ). **b**, The plots are ordered by functional group combinations: grasses (g); small herbs (sh); tall herbs (th); legumes (l). Symbols are means  $\pm 1$  standard error of the mean (s.e.m.) calculated from raw data.

and also tested whether such niche differentiation was greater for multi-functional group mixture types than for monofunctional group mixture types. Furthermore, we tested for character displacement by measuring relative differences in functional traits that reflect plant growth strategies<sup>16</sup>, height and specific leaf area (SLA) between species in mixtures (absolute difference between two species divided by the mean of the two).

Mixtures of mixture types had higher biomass than mixtures of monoculture types (Extended Data Table 1,  $P = 0.024$ ), and this pattern was consistent across functional group combinations. Net biodiversity effects and complementarity effects were larger for mixture types than for monoculture types (Fig. 2a, b,  $P = 0.017$  and  $P = 0.005$ , respectively; Table 1), indicating increased niche differentiation, with consistent results across the majority of functional group combinations, as illustrated in Fig. 2b. The positive effect of mixture types on net effects was strongest in grass mixtures, short-herb mixtures and mixtures of grasses with short herbs. Strong complementarity effects in response to selection history were found in these same mixtures, as well as in mixtures of legumes with

grasses or tall herbs. Positive sampling effects in response to selection history were found for grass mixtures, legume mixtures and mixtures of legumes with short herbs, but in general the response of sampling effects to selection history was slightly negative (Fig. 2a).

The stronger biodiversity effects obtained with mixture types in contrast to monoculture types were mirrored by larger functional trait differences between species in mixtures of mixture types rather than monoculture types. This was the case for relative height differences between species (Fig. 3a; Extended Data Table 2,  $P = 0.011$ ) and for relative differences in SLA (Fig. 3b,  $P < 0.001$ ). Mixture test communities with legumes showed particularly large SLA differences between mixture types (Extended Data Table 2,  $P < 0.001$ ). Functional diversity, calculated from height, SLA and additionally reproductive biomass, was greater in test communities of mixture types than of monoculture types (Fig. 3c and Extended Data Table 2). We found a trend of increased intraspecific functional diversity when monocultures were planted with monoculture types as compared with mixture types (mean difference =  $-0.202$ , paired  $t$ -test,  $P = 0.101$ ), indicating a broader spread of phenotypes within the monoculture type populations.

The association between increased biodiversity effects and increased functional trait differences of mixture types was reflected in marginally significant correlations across selection history treatments: complementarity effects rose with greater relative differences in SLA ( $P = 0.073$ ) and selection effects rose with greater relative differences in height ( $P = 0.074$ ). We expect the functional traits measured here to be representative of relevant niche dimensions. However, niche differentiation is probably multivariate, and additional traits such as rooting depth could be included in future experiments. Furthermore, such experiments should assess potential changes of traits during the course of an experiment; here we only measured them at the end.

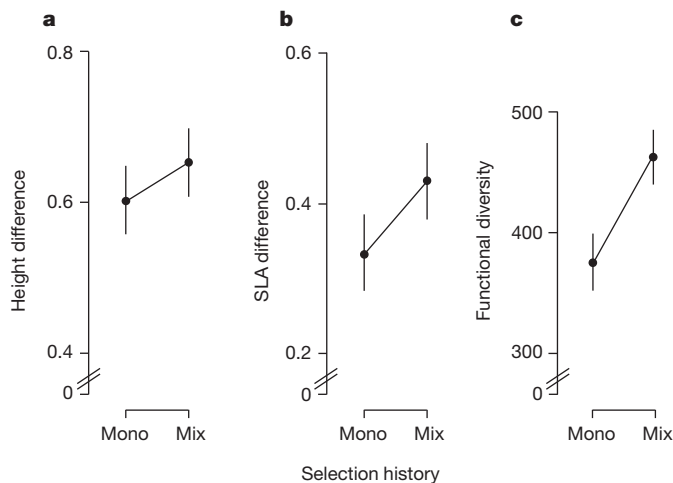
We demonstrate increased mixture performance and biodiversity effects in test communities of mixture types and provide evidence that these increases were driven by increased niche differentiation due to character displacement of functional traits between species. Our results were obtained with 12 typical grassland species of four different functional groups, supporting our hypothesis that increased biodiversity effects can result from selection for increased mixture performance and suggesting that these results may apply more generally. In a field-based extension of the current work, we found increased biodiversity effects across a wider range of species ( $n = 52$ ) and species richness levels (1, 2, 4, 8) for communities sharing a common selection history (Extended Data Fig. 2, interaction of selection history and species richness,  $P < 0.001$ ). Whereas mixtures may select for increased complementarity and character displacement between species, it is conceivable that selection pressures in monocultures select for greater defence against species-specific pathogens known to accumulate in monocultures<sup>8</sup>. We observed increased monoculture performance and reduced biodiversity effects in mixtures planted with monoculture types. Because mixture and monoculture types experienced selection environments for only 8 years, the standing variation at the beginning of the experiment may have already included genotypes or epigenetic variants pre-adapted for monoculture or mixture environments<sup>17,18</sup>.

**Table 1 | Net effect, complementarity effect and sampling effect on community biomass**

Source of variation	numDF	NE				CE			SE		
		denDF	F	P		denDF	F	P	denDF	F	P
Selection history											
Monoculture versus mixture	1	62.9	6.04	0.017		64.4	8.49	0.005	66.7	0.01	0.921
Functional group combination	9	61.4	2.4	0.021		61.1	1.05	0.410	64.7	1.91	0.067
Monoculture versus mixture $\times$ functional group combination	9	66.2	1.43	0.193		69.0	1.32	0.242	70.5	0.90	0.527
Random terms	<i>n</i>	VC	s.e.			VC	s.e.		VC	s.e.	
Monoculture versus mixture $\times$ species combination	86	0.942	0.415			0.123	0.086		0.117	0.056	
Residual	545	8.314	0.547			2.254	0.148		1.245	0.082	

Results of mixed-effects analysis of variance (ANOVA) for net effects (NE; untransformed), complementarity effects (CE; square-root transformed) and sampling effects (SE; square-root transformed). denDF, degrees of freedom of error term (which can be fractional in residual maximum likelihood analysis); numDF, degrees of freedom of term. F, variance ratio; *n*, number of replicates for random effects; P, error probability; s.e., standard error of variance component; VC, variance component.





**Figure 3 | Plants selected in mixtures show character displacement between species when grown in mixture.** a–c, After 8 years of selection in the Jena Experiment, mixture types (Mix) in comparison with monoculture types (Mono) showed character displacement between species. a, b, Relative differences between the two species in mixture for plant height ( $n = 219$  aggregated differences) (a) and SLA ( $n = 208$  aggregated differences) (b). c, Functional diversity<sup>30</sup> (calculated from height, SLA and reproductive biomass) of the two species in mixture (unit-less functional diversity index,  $n = 219$  aggregated values). Symbols are means  $\pm$  s.e.m. calculated from raw data. Means are averages over all 50 species combinations.

Niche differences between species decrease the strength of interspecific competition relative to intraspecific competition<sup>19</sup>. Thus, selection in high-diversity communities with high interspecific competition can be expected to favour genotypes with more distinct niches, reducing niche overlap and competition between species. Reduced interspecific competition could also result from the extension of the total community niche<sup>20</sup> in addition to, or instead of, finer division of currently used resources. The notion of diversity as a driver of plant population differentiation has been suggested in theory<sup>21</sup> and a field study has demonstrated that differential selection for monoculture and mixture types in grassland species can occur<sup>22</sup>, supporting the idea that local evolutionary changes arise from selection through competition<sup>23</sup> over short time scales<sup>24</sup>. In addition, species diversity has been shown to influence plant traits associated with light and resource uptake such as shoot, leaf and stem length<sup>25,26</sup>, with increasing functional diversity within communities<sup>27</sup>. In our study, mixture types across all functional groups showed greater relative differences in height and SLA between species in mixtures. These between-species differences may have resulted from directional selection or selection for increased plasticity in mixture types. Finally, infrared spectral fingerprints obtained for 8 of the 12 species showed significant differences in metabolic profiles between monoculture- and mixture-type individuals (Extended Data Fig. 3), reflecting differential chemistry.

We demonstrated an interaction between selection community and test community diversity, with increased community performance of mixture types. Furthermore, mixture types exhibited interspecific trait divergence, which potentially explains the increased biodiversity effects for mixtures of mixture types in our study. The consequences of selection in long-term field experiments have previously been considered in theory<sup>28</sup>, and selection has been shown to drive community dynamics in microcosms<sup>29</sup>; our results demonstrate that such evolutionary processes can cause the emergence of stronger biodiversity effects over time in plant biodiversity experiments. Capturing the potential of this production-enhancing niche differentiation in diverse communities may have profound impacts for agricultural and forestry applications, for example by achieving increased productivity in crop mixtures using varieties that have been selected in diverse planting regimes. This novel mechanism also implies that species losses can affect not only ecosystem

functioning in the short term, but also the long-term trajectory of biological communities.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 11 June 2014; accepted 17 September 2014.

Published online 15 October 2014.

1. Tilman, D., Reich, P. B. & Knops, J. M. H. Biodiversity and ecosystem stability in a decade-long grassland experiment. *Nature* **441**, 629–632 (2006).
2. Reich, P. B. *et al.* Impacts of biodiversity loss escalate through time as redundancy fades. *Science* **336**, 589–592 (2012).
3. Hector, A. *et al.* General stabilizing effects of plant diversity on grassland productivity through population asynchrony and overyielding. *Ecology* **91**, 2213–2220 (2010).
4. Kulmatiski, A., Beard, K. H. & Heavilin, J. Plant–soil feedbacks provide an additional explanation for diversity–productivity relationships. *Proc. R. Soc. B* **279**, 3020–3026 (2012).
5. Roscher, C., Thein, S., Schmid, B. & Scherer-Lorenzen, M. Complementary nitrogen use among potentially dominant species in a biodiversity experiment varies between two years. *J. Ecol.* **96**, 477–488 (2008).
6. Mueller, K. E., Tilman, D., Fornara, D. A. & Hobbie, S. E. Root depth distribution and the diversity–productivity relationship in a long-term grassland experiment. *Ecology* **94**, 787–793 (2013).
7. Fornara, D. A. & Tilman, D. Plant functional composition influences rates of soil carbon and nitrogen accumulation. *J. Ecol.* **96**, 314–322 (2008).
8. Janzen, D. H. Herbivores and the number of tree species in tropical forests. *Am. Nat.* **104**, 501–508 (1970).
9. Connell, J. H. in *Dynamics of Populations* (eds den Boer, P. J. & Gradwell, G. R.) 298–312 (Center for Agricultural Publishing and Documentation, 1971).
10. Dayan, T. & Simberloff, D. Ecological and community-wide character displacement: the next generation. *Ecol. Lett.* **8**, 875–894 (2005).
11. Wacker, L., Baudois, O., Eichenberger-Glinz, S. & Schmid, B. Effects of plant species richness on stand structure and productivity. *J. Plant Ecol.* **2**, 95–106 (2009).
12. Zeller, S. L., Kalinina, O., Flynn, D. F. B. & Schmid, B. Mixtures of genetically modified wheat lines outperform monocultures. *Ecol. Appl.* **22**, 1817–1826 (2012).
13. Tilman, D., Cassman, K. G., Matson, P. A., Naylor, R. & Polasky, S. Agricultural sustainability and intensive production practices. *Nature* **418**, 671–677 (2002).
14. Roscher, C. *et al.* The role of biodiversity for element cycling and trophic interactions: an experimental approach in a grassland community. *Basic Appl. Ecol.* **5**, 107–121 (2004).
15. Loreau, M. & Hector, A. Partitioning selection and complementarity in biodiversity experiments. *Nature* **412**, 72–76 (2001).
16. Westoby, M., Falster, D. S., Moles, A. T., Vesk, P. A. & Wright, I. J. Plant ecological strategies: some leading dimensions of variation between species. *Annu. Rev. Ecol. Syst.* **33**, 125–159 (2002).
17. Turkington, R. & Harper, J. L. Growth, distribution and neighbor relationships of *Trifolium repens* in a permanent pasture. 2. Interspecific and intraspecific contact. *J. Ecol.* **67**, 219–230 (1979).
18. Fakheran, S. *et al.* Adaptation and extinction in experimentally fragmented landscapes. *Proc. Natl Acad. Sci. USA* **107**, 19120–19125 (2010).
19. Chesson, P. Mechanisms of maintenance of species diversity. *Annu. Rev. Ecol. Syst.* **31**, 343–366 (2000).
20. Salles, J. F., Poly, F., Schmid, B. & Le Roux, X. Community niche predicts the functioning of denitrifying bacterial assemblages. *Ecology* **90**, 3324–3332 (2009).
21. Vellend, M. & Geber, M. A. Connections between species diversity and genetic diversity. *Ecol. Lett.* **8**, 767–781 (2005).
22. Lipowsky, A., Schmid, B. & Roscher, C. Selection for monoculture and mixture genotypes in a biodiversity experiment. *Basic Appl. Ecol.* **12**, 360–371 (2011).
23. Taylor, D. R. & Aarssen, L. W. Complex competitive relationships among genotypes of 3 perennial grasses: implications for species coexistence. *Am. Nat.* **136**, 305–327 (1990).
24. Thompson, J. N. Rapid evolution as an ecological process. *Trends Ecol. Evol.* **13**, 329–332 (1998).
25. Gubsch, M. *et al.* Differential effects of plant diversity on functional trait variation of grass species. *Ann. Bot.* **107**, 157–169 (2011).
26. Roscher, C., Schmid, B., Buchmann, N., Weigelt, A. & Schulze, E. D. Legume species differ in the responses of their functional traits to plant diversity. *Oecologia* **165**, 437–452 (2011).
27. Roscher, C. *et al.* A functional trait-based approach to understand community assembly and diversity–productivity relationships over 7 years in experimental grasslands. *Perspect. Plant Ecol. Evol. Syst.* **15**, 139–149 (2013).
28. Strauss, S. Y., Lau, J. A., Schoener, T. W. & Tiffin, P. Evolution in ecological field experiments: implications for effect size. *Ecol. Lett.* **11**, 199–207 (2008).
29. Hansen, S. K., Rainey, P. B., Haagen, J. A. J. & Molin, S. Evolution of species interactions in a biofilm community. *Nature* **445**, 533–536 (2007).
30. Petchey, O. L. & Gaston, K. J. Functional diversity (FD), species richness and community composition. *Ecol. Lett.* **5**, 402–411 (2002).

**Acknowledgements** This study was supported by the Swiss National Science Foundation (grant number 130720 to B.S.) and the University Research Priority

Program Global Change and Biodiversity of the University of Zurich. Thanks to D. Trujillo Villegas, L. Oesch, T. Zwimpfer, M. Furler, R. Husi, the gardeners of the Jena Experiment and student helpers for technical assistance. G.B.D.D. acknowledges the NWO-ALW VIDI grant scheme for financial support.

**Author Contributions** B.S. and J.S.P. conceptualized the study; D.Z.-D. designed the experimental procedure and carried out the experiment with the help of B.S., D.F.B.F. and V.Y.; B.S., D.Z.-D. and D.F.B.F. analysed the data; D.Z.-D., B.S. and D.F.B.F. wrote the

paper with input from J.S.P., V.Y. and G.B.D.D. All authors discussed study design, field and glasshouse work, and analysis.

**Author Information** Data have been deposited at the Dryad Data Repository (<http://dx.doi.org/10.5061/dryad.750df>). Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to B.S. ([bernhard.schmid@ieu.uzh.ch](mailto:bernhard.schmid@ieu.uzh.ch)) or D.F.B.F. ([dan.flynn@ieu.uzh.ch](mailto:dan.flynn@ieu.uzh.ch)).

## METHODS

**Experimental setup.** To test whether plant types selected over 8 years in mixtures outperform those types selected in monocultures when assembled in mixture test communities, and vice versa for types selected in monocultures, we first selected 12 of 16 species grown in large monoculture and mixture plots in the Jena Experiment, Germany (50°55' N, 11°35' E, 130 m above sea level) (<http://www.the-jena-experiment.de>; see ref. 14 for experimental details). Altogether there are 60 species in the Jena Experiment, but 44 of them do not occur in large monoculture plots. Three species from each of the following four functional groups of plant species used in the Jena Experiment were selected (a fourth species in each functional group was omitted because of poor growth in monoculture): grasses (*Festuca pratensis*, *Festuca rubra*, *Poa pratensis*), small herbs (*Plantago lanceolata*, *Prunella vulgaris*, *Veronica chamaedrys*), tall herbs (*Crepis biennis*, *Galium mollugo*, *Geranium pratense*) and legumes (*Lathyrus pratensis*, *Onobrychis viciifolia*, *Trifolium repens*) (Extended Data Fig. 1). These species had an 8-year community history growing in monocultures or mixtures consisting of plants belonging to a single functional group or to all four functional groups. In April 2010, we collected 4,900 plant cuttings from 48 of the 82 plots in the Jena Experiment.

We used these plant cuttings to establish plots in an experimental garden in slug enclosure compartments at the University of Zurich, Switzerland (47°23' N, 8°33' E, 534 m above sea level) with an identical plant composition to the plots in Jena from which the cuttings were collected. We added a layer of potting soil (BF 4, De Baat; Extended Data Table 3) to the soil in each plot to make sure the plants established. Netting around each plot minimized the possibility of cross-pollination between the same species from different community histories. The cuttings were used for the propagation of further cuttings or to produce seeds for our study.

Using 25 of the plant cuttings of each species from the three field community histories, we generated further cuttings, from April 2010 until September 2010, in pots in the experimental glasshouse to set up the first block of the experiment. Finally, the cuttings used in the experiment were the result of six rounds of propagation. This was done to reduce potential carry-over effects. Each set of cuttings was timed according to the slowest growing species, every 3 to 4 weeks.

During the summer of 2010, we collected seed material from the experimental garden plots for the second block of this experiment. The seed material was dried in a glasshouse compartment. We cleaned the seeds from the husks/pods and stored them at 10–15 °C, 50% humidity in a climate chamber. Once all the seeds were cleaned, they were treated with cold stratification at 5 °C for 2 months. Seeds were germinated in a 10.5-h day regime with 14–19 °C day and 10–16 °C night temperature.

In November 2010 we transplanted randomly selected individuals that were cuttings from the 25 original cuttings. We planted monocultures of four plants or two-species mixtures of two plus two plants into pots (4,275 cm<sup>3</sup>) filled with neutral agricultural soil (Extended Data Table 3) according to a diallel design containing all possible combinations of species within and among functional groups according to available plant material (Extended Data Fig. 1). Cuttings of the legume *Onobrychis viciifolia* were not successfully propagated and were therefore excluded from the first block. In October 2011, we transplanted seedlings into pots following the same procedure and design. In total we planted 12 monoculture and 50 two-species combinations as test communities with plants of three types of selection history: monoculture types and mixture types taken from mixtures of a single functional group (monofunctional group mixture types) and from mixtures of four different plant functional groups (multifunctional group mixture types), replicated, if possible, three times for cuttings and three times for seedlings ( $n = 855$  pots; Extended Data Fig. 1). Single pots always contained four plants of a single selection history. Plant traits and biomass were measured 20 weeks after planting for the block established with cuttings and the block established with seedlings. To exclude effects of plant–soil feedbacks, we used a neutral growth substrate (50% agricultural sugarbeet soil, 25% sand, 25% Perlite; Ricoter AG; Extended Data Table 3) throughout the experiment.

Once the plants were transplanted into the pots, glasshouse conditions were set to natural summer day length and day temperatures of 20 °C and night temperatures

of 17 °C. To supplement sunlight, additional light was provided at a maximum of 30 kLux (Metallhalogenlamps 400 W, Iwasaki MT 400 DL/BH). Shading was at 20 kLux. To compensate for overheating, an adiabatic cooling system (Airwatech) was used. The plants were watered in the trays to make sure that each individual received equal water volume. Seedlings that died in the first 2 weeks were replaced with seedlings of the same age. Pot locations were randomized in the glasshouse without reference to history or species combination.

**Measurements and harvest.** Height and leaf number were measured at planting to ensure that the cuttings and seedlings were standardized. After 20 weeks of growth in the pots, plant height was measured again and the aboveground biomass of each individual was harvested at ground level. The inflorescence, if present at harvest, was collected separately and the dry biomass weighed as an indication of reproductive effort. SLA of representative leaves of each species in a pot was measured by scanning fresh leaves (Licor LI-3100) immediately after harvest and determining the mass of the same leaves after drying. Research assistants assisted in the regular measurements and harvesting of plants at the end of the experiment, and these assistants were not informed of the specific experimental treatments.

**Statistical analysis.** We compared the performance of plants in monoculture versus mixture test communities after being selected in either monoculture or mixture communities for 8 years using the mean aboveground dry biomass of our test communities as the response variable. We used general mixed models using residual maximum likelihood (REML) and summarized results in ANOVA tables. Significance tests were based on approximate  $F$ -tests using appropriate error terms and denominator degrees of freedom. The fixed terms in the models were: block (cuttings versus seedlings); planted diversity (monoculture versus mixture); planted functional group diversity (monoculture versus monofunctional group mixture versus multifunctional group mixture); selection history (monoculture selection versus mixture selection, resulting in monoculture versus mixture types); functional group diversity of the selection history (monoculture selection versus monofunctional group selection versus multifunctional group selection); presence or absence of legumes in the test community; functional group combination within a test community; and interactions among these. Glasshouse table, pot within table, and species combination were used as random terms. No outliers were excluded from this analysis; the full model is presented in Extended Data Table 1.

To assess biodiversity effects we followed the additive partitioning method previously described<sup>15</sup>, and partitioned the net effect (NE) into complementarity effects (CE) and sampling effects (SE). Calculations were based on the difference between the observed yield of each species in the mixture and mean monoculture yield for that species in the corresponding block and for that specific selection history. Absolute values of CE and SE were square-root transformed and the original signs put back on the transformed values for analysis<sup>15</sup>. There were 545 pots with mixture test communities for which NE, CE and SE could be calculated. For the analysis of the biodiversity effects we could simplify the full mixed model described earlier to a concise model presented in Table 1.

Relative differences in height and SLA (absolute difference between two species divided by the mean of the two), as well as functional diversity (FD), were calculated between species in mixtures. Differences in these measures between mixtures of monoculture versus mixture types might be associated with the differences in biodiversity effects between mixtures of monoculture versus mixture type; this would identify character displacement as a potential mechanism underlying positive biodiversity effects. FD was calculated following ref. 30 using height, SLA and reproductive biomass as functional traits. Relative differences and FD values between species means in mixture test communities were aggregated at the level of species combination by selection history by block ( $n = 219$ ), using means calculated for each species by selection history by block combination. Statistical analysis was done with the full model described earlier and is presented in Extended Data Table 2.

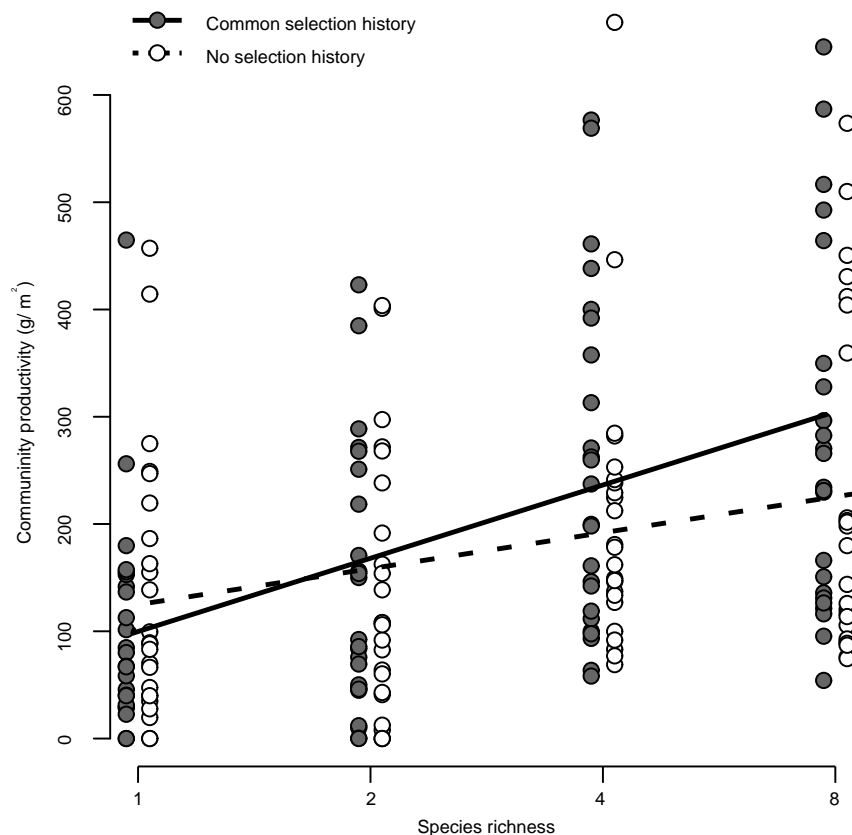
Statistical analyses were conducted using the software products R, version 2.15.3 (R Development Core Team), and GenStat, version 16 (VSN International).



		Selection history											
		Monoculture				Monofunctional group				Mix-functional group			
		Grass			Short Herb			Tall Herb			Grass		
		g1	g2	g3	sh1	sh2	sh3	th1	th2	th3	g1	g2	g3
Planted diversity	Monoculture	g1	3								3		
		g2		3								3	
		g3			3								3
		sh1				3							
		sh2					3						
		sh3						3					
		th1							3				
		th2								3			
		th3											3
	Monofunctional mixture	l1											
		l2											
		l3											
		g1			3								
		g2				3							
		g3					3						
		sh1						3					
		sh2							3				
		sh3								3			
	Mix-functional mixture	th1											
		th2											
		th3											
		l1											
		l2											
		l3											
		g1				3							
		g2					3						
		g3						3					
	Monoculture	sh1											
		sh2											
		sh3											
		th1											
		th2											
		th3											
		l1											
		l2											
		l3											
	Monofunctional mixture	g1											
		g2											
		g3											
		sh1											
		sh2											
		sh3											
		th1											
		th2											
		th3											
	Mix-functional mixture	l1											
		l2											
		l3											
		g1											
		g2											
		g3											
		sh1											
		sh2											
		sh3											
	Monoculture	th1											
		th2											
		th3											
		l1											
		l2											
		l3											
		g1											
		g2											
		g3											
	Monofunctional mixture	sh1											
		sh2											
		sh3											
		th1											
		th2											
		th3											
		l1											
		l2											
		l3											
	Mix-functional mixture	g1											
		g2											
		g3											
		sh1											
		sh2											
		sh3											
		th1											
		th2											
		th3											
	Monoculture	l1											
		l2											
		l3											
		g1											
		g2											
		g3											
		sh1											
		sh2											
		sh3											
	Monofunctional mixture	th1											
		th2											
		th3											
		l1											
		l2											
		l3											
		g1											
		g2											
		g3											
	Mix-functional mixture	sh1											
		sh2											
		sh3											
		th1											
		th2											
		th3											
		l1											
		l2											
		l3											

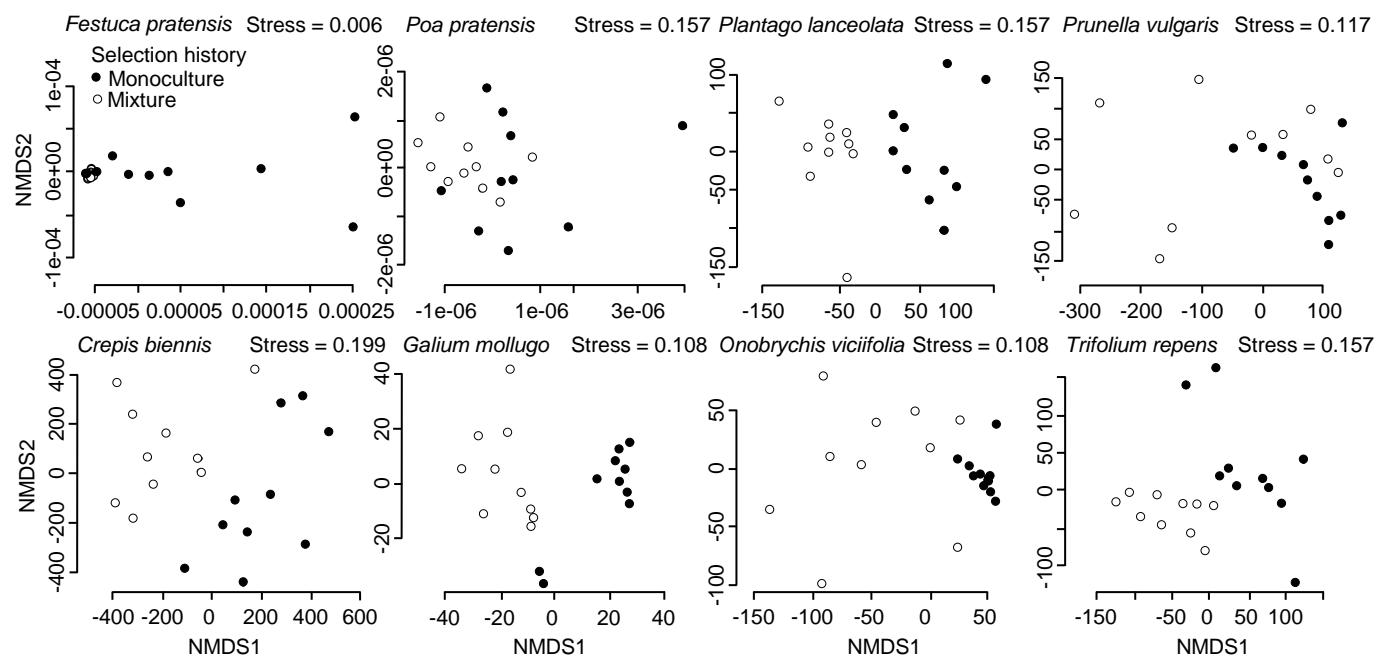
**Extended Data Figure 1 | Designed number of pots planted for each combination of species.** Plants from three different selection histories (monoculture, plot containing one species; monofunctional group, plot containing at least four species of the same functional group of plants; mixed-functional group, plot containing at least four species of four different functional groups) were grown in three different types of test communities (monoculture, monofunctional mixture, mixed-functional mixture). Twelve species in the Jena Experiment were chosen from all four functional groups: grass (g) (*Festuca pratensis*, *Festuca rubra*, *Poa pratensis*), small herb (sh) (*Plantago lanceolata*, *Prunella vulgaris*, *Veronica chamaedrys*), tall herb (th) (*Crepis biennis*, *Galium mollugo*, *Geranium pratense*), legume (l) (*Lathyrus pratensis*, *Onobrychis viciifolia*, *Trifolium repens*); numbers after the letter abbreviations refer to the different species. This design was used once with plants raised from cuttings (Block 1) and once with plants raised from seedlings

(Block 2). Overall we aimed to obtain the same 12 monocultures and 48 two-species combinations as test communities for each block. Availability of species precluded some of the two-species combinations in each block, such that they had to be replaced by other combinations. This yielded a total of 50 combinations across the two blocks, with several that were unique within a block. Each monoculture and each two-species combination was assembled three times for each of the three types of selection histories in each block. Some monocultures and some two-species combinations could not be realized with all types of selection histories in both blocks. Overall, there were 855 pots, 168 monocultures and 687 two-species mixtures; for 545 of the latter, the net biodiversity effect could be partitioned into complementarity and sampling effects. Some missing monocultures precluded the calculation of biodiversity effects in certain mixtures.



**Extended Data Figure 2 | Biodiversity-productivity relationship is stronger for plants with a common selection history.** Aboveground net primary productivity of communities in an experimental manipulation of plant species richness and selection history (common history versus no common history). In this experiment, species represented an expanded set from the present experiment (52 species), and were planted within a large-scale field experiment in Jena, Germany, on mixed soil from 48 plots from which plants had been selected, thus equalizing potential effects of soil legacy among treatments. Plants without selection history were grown from seed from a seed company,

while plants with selection history were seed progeny from plots of exactly the same species composition as the one in which they were replanted (same propagation procedure as for the 12 species used in the test communities of the present study). The slope of the biodiversity-productivity relationship was steeper for plants with a common selection history (significance of slope differences tested with interaction term  $\log(\text{species richness}) \times \text{selection history}$  in mixed model with random-effects factor for 48 specific plant communities;  $P < 0.001$ ,  $n = 96$ ).



**Extended Data Figure 3 | Selection for different biochemical features in monocultures and mixtures.** Ordinations (non-metric multidimensional scaling (NMDS)) of second derivative of spectral wavenumbers of 8 of the 12 species used in the present study, showing effects of 8-year selection history on plant individuals derived from monoculture and mixture communities (Jena

Experiment). This can be an indication of selection for different biochemical features over 8 years in monoculture and mixtures. Stress values reflect a measure of goodness of fit for NMDS, with lower values showing better representation of the original data.



**Extended Data Table 1 | Results of mixed-effects ANOVA for the aboveground biomass of test communities 20 weeks after transplanting plants into pots**

Source of variation	numDf	denDf	<i>F</i>	<i>P</i>
Seedlings versus cuttings	1	65.2	5.89	<b>0.018</b>
Selection history: monoculture versus mixture	1	717.4	0.55	0.457
Selection history: monofunctional group versus multi-functional group mixture	1	707.1	0.11	0.735
Planted diversity: monoculture versus mixture	1	41.2	4.84	<b>0.034</b>
Planted diversity: monofunctional group versus multi-functional group mixture	1	41.9	0.09	0.767
Selection history: monoculture versus mixture × planted diversity: monoculture versus mixture	1	714.3	5.15	<b>0.024</b>
Selection history: monofunctional group versus multi-functional group mixture × planted diversity: monofunctional group versus multi-functional group mixture	3	713.4	1.30	0.274
Legumes	1	44.5	38.02	<b>&lt;0.001</b>
Functional group combination	7	43.8	2.51	<b>0.029</b>
Selection history: monoculture versus mixture × legumes	1	727.7	4.04	<b>0.045</b>
Planted diversity: monoculture versus mixture × legumes	1	46.3	0.14	0.712
Selection history: monoculture versus mixture × functional group combination	7	729.2	1.17	0.318
Planted diversity: monoculture versus mixture × functional group combination	2	42.4	0.41	0.664
Selection history: monoculture versus mixture × planted diversity: monoculture versus mixture × legumes	1	704.0	1.27	0.260
Selection history: monoculture versus mixture × planted diversity: monoculture versus mixture × functional group combination	2	731.1	1.58	0.207
Selection history : monofunctional group versus multi-functional group mixture × planted diversity: monofunctional group versus multi-functional group mixture × legumes	3	725.8	0.45	0.719
Selection history : monofunctional group versus multi-functional group mixture × planted diversity: monofunctional group versus multi-functional group mixture × functional group combination	8	719.6	2.23	<b>0.023</b>
Block × selection history: monoculture versus mixture	1	729.9	2.9	0.089
Block × planted diversity: monoculture versus mixture	1	50.5	0.00	0.986
Block × selection history: monoculture versus mixture × planted diversity: monoculture versus mixture	1	722.1	2.08	0.149
Block × selection history: monofunctional group versus multi-functional group mixture × planted diversity: monofunctional group versus multi-functional group mixture	5	361.9	0.68	0.642
Random terms	<i>n</i>	VC	s.e.	
Block × glasshouse table	44	0.662	0.235	
Species combination	62	0.772	0.764	
Block × species combination	109	2.905	0.822	
Residual	855	6.708	0.367	

denDf, degrees of freedom of error term (which can be fractional in residual maximum likelihood analysis); numDf, degrees of freedom of term. *F*, variance ratio; *n*, number of replicates for random effects; *P*, error probability; s.e., standard error of variance component; VC: variance component.

**Extended Data Table 2 | Results of mixed-effects ANOVA for relative height difference, relative SLA difference and for functional diversity**

Source of variation	Height difference				SLA difference				Functional diversity			
	numDf	denDf	F	P	numDf	denDf	F	P	numDf	denDf	F	P
Seedlings versus cuttings	1	36.1	0.10	0.757	1	28.6	26.21	<b>&lt;0.001</b>	1	33.7	3.07	0.089
Selection history: monoculture versus mixture	1	30.7	7.25	<b>0.011</b>	1	27.2	20.62	<b>&lt;0.001</b>	1	32.3	17.42	<b>&lt;0.001</b>
Planted diversity: monofunctional group versus multi-functional group mixture	-	-	-	-	1	38.0	1.24	0.272	1	37.8	4.23	<b>0.047</b>
Functional group combinations with legumes	1	37.1	8.20	<b>0.007</b>	1	38.7	0.28	0.602	1	39.2	3.69	<b>0.062</b>
Functional group combinations rest	8	36.8	0.90	0.525	7	38.5	3.25	<b>0.008</b>	7	38.2	4.64	<b>&lt;0.001</b>
Seedlings versus cuttings x selection history: monoculture versus mixture	1	83.4	37.24	<b>&lt;0.001</b>	1	82.3	91.76	<b>&lt;0.001</b>	1	85.3	36.41	<b>&lt;0.001</b>
Seedlings versus cuttings x planted diversity: monofunctional group versus multi-functional group mixture	-	-	-	-	1	27.7	4.34	<b>0.047</b>	1	33.3	2.84	0.101
Seedlings versus cuttings x functional group combinations with legumes	1	36.4	0.43	0.517	-	-	-	-	1	34.6	5.97	<b>0.02</b>
Seedlings versus cuttings x functional group combinations rest	8	36.1	0.55	0.809	-	-	-	-	-	-	-	-
Seedlings versus cuttings x functional group combinations all	-	-	-	-	8	29.2	1.20	0.334	-	-	-	-
Selection history: monoculture versus mixture x planted diversity: monofunctional group versus multi-functional group mixture	-	-	-	-	-	-	-	-	1	32.3	1.37	<b>0.251</b>
Selection history: monoculture versus mixture x functional group combinations with legumes	1	31.1	2.11	0.156	1	27.6	18.13	<b>&lt;0.001</b>	1	32.7	13.00	<b>0.001</b>
Selection history: monoculture versus mixture x functional group combinations rest	8	30.7	2.48	<b>0.034</b>	8	27.2	0.60	0.773	-	-	-	-
Seedlings versus cuttings x selection history: monoculture versus mixture x planted diversity: monofunctional group versus multi-functional group mixture	-	-	-	-	-	-	-	-	1	88	12.08	<b>&lt;0.001</b>
Seedlings versus cuttings x selection history: monoculture versus mixture x functional group combinations all	-	-	-	-	9	83.7	8.96	<b>&lt;0.001</b>	-	-	-	-
Seedlings versus cuttings x selection history: monoculture versus mixture x functional group combinations with legumes	1	84.9	20.11	<b>&lt;0.001</b>	-	-	-	-	1	85.8	5.41	<b>0.022</b>
Seedlings versus cuttings x selection history: monoculture versus mixture x functional group combinations rest	8	84.6	3.22	<b>0.003</b>	-	-	-	-	21	92.3	2.51	<b>0.001</b>
Random terms	<i>n</i>	VC	s.e.		<i>n</i>	VC	s.e.		<i>n</i>	VC	s.e.	
Species combination	50	0.006	0.031		50	0.040	0.015		50	4139	4449	
Selection history: monoculture versus mixture x species combination	93	0.005	0.002		89	0.007	0.003		93	6784	2010	
Seedlings versus cuttings x species combination	88	0.160	0.041		85	0.025	0.008		88	14483	4023	
Residual	219	0.002	0.003		208	0.005	0.001		219	2152	350	

denDf, degrees of freedom of error term (which can be fractional in residual maximum likelihood analysis); numDf, degrees of freedom of term. *F*, variance ratio; *n*, number of replicates for random effects; *P*, error probability; s.e., standard error of variance component; VC: variance component.

Extended Data Table 3 | Composition of the experimental substrate

Parameter	Unit	GVZ	AGR
Carbon*	µg/g	411.5	34.60
Hydrogen*	µg/g	47.7	4.6
Nitrogen*	µg/g	8.0	2.1
pH <sup>†</sup>		5.4	7.9
Organic matter <sup>†</sup>	%(mass)	55	3.5
Clay <sup>†</sup>	%(mass)	1.0	1.0
Silt <sup>†</sup>	%(mass)	1.0	1.0
Nitrate <sup>†</sup>	mg/l	439	730
Ammonium <sup>†</sup>	mg/l	0.7	2.5
Phosphorus <sup>†</sup>	mg/l	20	0.3
Potassium <sup>†</sup>	mg/l	54	127
Calcium <sup>†</sup>	mg/l	119	187
Magnesium <sup>†</sup>	mg/l	43	40

Composition of 1 g of substrate GVZ Tref GO PP 7000 (BF4: black peat; white peat; clay; mineral fertilizer, 1.3 kg m<sup>-3</sup>) and neutral agricultural soil (50% sugarbeet soil, sieved; 25% washed river sand, 0–2 mm; 25% perlite, 2–6 mm; AGR; RicoterAG). All units in mg l<sup>-1</sup> are per litre extract solution.

\* Composition determined using elemental analysis.

† Composition determined by Ibu (Laboratory for Soil Analysis, Thun, Switzerland), program 40 analysis.



# Nodal signalling determines biradial asymmetry in *Hydra*

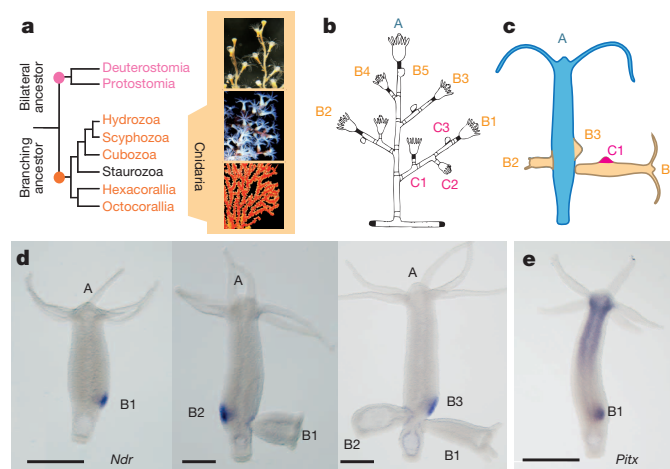
Hiroshi Watanabe<sup>1</sup>, Heiko A. Schmidt<sup>2</sup>, Anne Kuhn<sup>1</sup>, Stefanie K. Höger<sup>1</sup>, Yigit Kocagöz<sup>1</sup>, Nico Laumann-Lipp<sup>1</sup>, Suat Özbek<sup>1</sup> & Thomas W. Holstein<sup>1</sup>

In bilaterians, three orthogonal body axes define the animal form, with distinct anterior–posterior, dorsal–ventral and left–right asymmetries. The key signalling factors are Wnt family proteins for the anterior–posterior axis, Bmp family proteins for the dorsal–ventral axis and Nodal for the left–right axis<sup>1</sup>. Cnidarians, the sister group to bilaterians, are characterized by one oral–aboral body axis, which exhibits a distinct biradiality of unknown molecular nature. Here we analysed the biradial growth pattern in the radially symmetrical cnidarian polyp *Hydra*, and we report evidence of Nodal in a pre-bilaterian clade. We identified a *Nodal-related* gene (*Ndr*) in *Hydra magnipapillata*, and this gene is essential for setting up an axial asymmetry along the main body axis. This asymmetry defines a lateral signalling centre, inducing a new body axis of a budding polyp orthogonal to the mother polyp's axis. *Ndr* is expressed exclusively in the lateral bud anlage and induces *Pitx*, which encodes an evolutionarily conserved transcription factor that functions downstream of Nodal. Reminiscent of its function in vertebrates<sup>2,3</sup>, Nodal acts downstream of  $\beta$ -Catenin signalling. Our data support an evolutionary scenario in which a 'core-signalling cassette' consisting of  $\beta$ -Catenin, Nodal and *Pitx* pre-dated the cnidarian–bilaterian split. We presume that this cassette was co-opted for various modes of axial patterning: for example, for lateral branching in cnidarians and left–right patterning in bilaterians.

An animal body plan can be described by using the axes of a Cartesian coordinate system<sup>1,4</sup>. Bilateral animals, for example, exhibit one major body axis, passing from anterior to posterior, a horizontal plane of symmetry separating the dorsal and ventral sides and a mid-sagittal plane of symmetry separating the left and right sides. How these body axes and bilaterality arose in ancestral animals, which have a simpler body plan, is unknown<sup>1</sup>. An answer to this question might be found in the sister group to the bilaterian clade, cnidarians: this diverse clade is more than 600 million years old and contains jellyfish (Scyphozoa, Staurozoa and in Hydrozoa), corals (in Octocorallia and in Hexacorallia), the freshwater polyp *Hydra* (in Hydrozoa) and less familiar forms such as sea anemones (in Hexacorallia) and hydroids (in Hydrozoa)<sup>5</sup> (Fig. 1a). The essence of the simple, gastrula-shaped cnidarian body plan is radiality with a major oral–aboral body axis<sup>5,6</sup>. Perfect radial symmetry is, however, uncommon, since most cnidarians exhibit morphological features of distinct biradiality either in their internal organization (anthozoans) and/or in colonial branching patterns (anthozoans and hydrozoans) (Fig. 1b). Even in the freshwater polyp *Hydra*, whose morphology shows perfect radial symmetry, evidence for biradiality has been reported during the formation of buds, which arise in a biradial pattern<sup>7</sup> (Fig. 1c), similar to branching colonial hydrozoans (Fig. 1b).

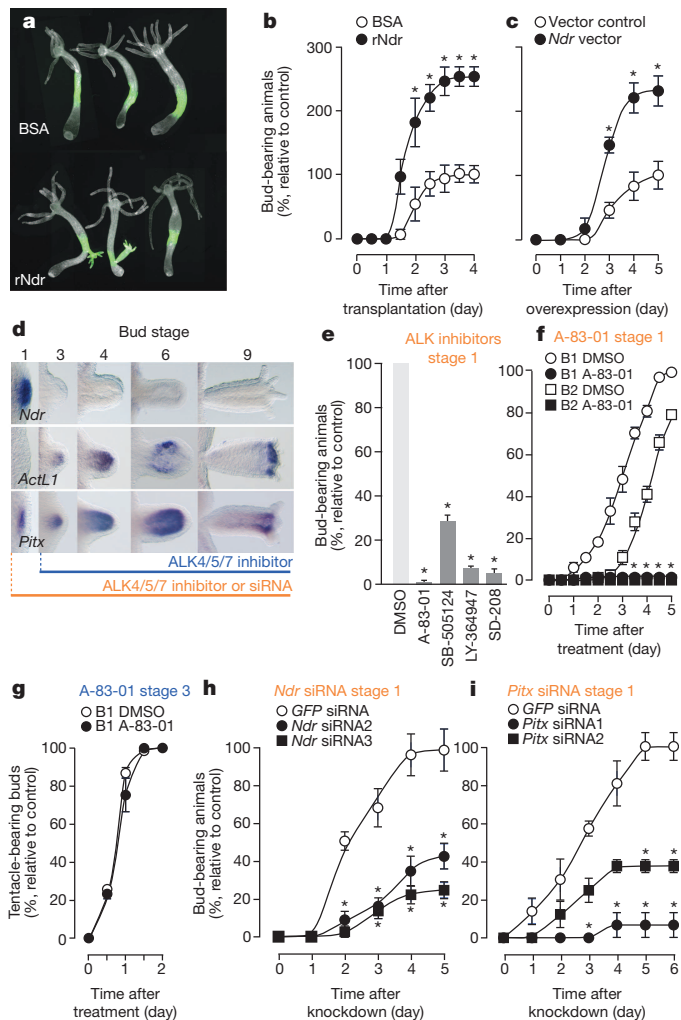
To identify the molecular factors causing biradial asymmetry in cnidarians, we systematically searched in cnidarian genome databases for members of the Tgf- $\beta$  superfamily known to be involved in defining the dorsal–ventral or left–right axis in bilaterians. Based on molecular phylogenetic analyses and experimental results, sequences were grouped into four categories: Activin and Tgf- $\beta$ , Bmp2/4, Bmp5–8 and Nodal

families (where cnidarian Bmp2/4, for example, is the common ancestor of bilaterian Bmp2 and Bmp4) (Extended Data Figs 1–3 and Supplementary Table 1). Of particular interest was a *Nodal-related* gene (hereafter denoted *Ndr*), which tends to cluster with bilaterian *Nodal* (Extended Data Fig. 2). Our phylogenomic analysis revealed that antagonists of Nodal signalling (that is, members of the Cerberus and Dan family), the signal transduction molecule Smad2/3, and the downstream target transcription factors *Pitx*, *FoxA* and *Lhx1/5* are encoded in cnidarian genomes (Extended Data Figs 1 and 4–6 and Supplementary Table 1). This finding suggests that in the last ancestor of Cnidaria and Bilateria, some of these components of the Nodal-signalling pathway had already been specifically deployed. Among the Tgf- $\beta$  family members encoded by *Hydra*, only *H. magnipapillata Ndr* exhibited a clear asymmetrical expression pattern in the presumptive budding zone and in early buds, and when the buds developed further, its expression disappeared (Fig. 1d). By comparison, the *Hydra* genes orthologous to



**Figure 1 | Branching morphology in cnidarians and expression of Nodal-signalling genes in *Hydra*.** **a**, A simplified metazoan tree with the major cnidarian clades exhibiting a branching (orange text) or non-branching (black text) morphology. The images show *Obelia dichotoma* (Hydrozoa) (top), a *Teleso* sp. (Octocorallia) (centre) and *Melithaea flabellifera* (Octocorallia) (bottom). (Images courtesy of H. Namikawa (top and bottom) and Y. Imahara (centre).) **b**, A colony of an aetheate hydrozoan shows a branching pattern with consecutive polyp generations (A, B and C) that remain attached (from ref. 25). **c**, The biradial pattern in *Hydra* bud formation<sup>7</sup> is related to the colonial branching pattern (**b**). A mother polyp (A; blue) forms continuous buds (B1–B3; yellow) that arise perpendicularly to the oral–aboral axis in the budding zone and detach. The buds are induced directly opposite each other, resulting in an alternating pattern of lateral branching. C1 (pink) indicates the position of the third bud generation. (Unpublished image courtesy of B. Hobmayer.) **d**, **e**, The biradial expression patterns of *H. magnipapillata Ndr* (**d**; blue) and *Pitx* (**e**; blue) in the bud anlage of budless and budding *Hydra* polyps with emerging buds (B1–B3). Scale bar, 2 mm.

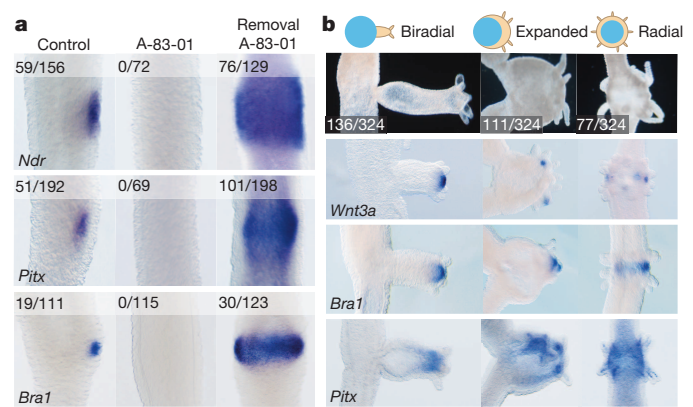
<sup>1</sup>Department of Molecular Evolution and Genomics, Centre for Organismal Studies (COS), Heidelberg University, 69120 Heidelberg, Germany. <sup>2</sup>Center for Integrative Bioinformatics Vienna (CIBIV), Max F. Perutz Laboratories (MFPL), University of Vienna/Medical University of Vienna, 1030 Vienna, Austria.



**Figure 2 | Requirement for Nodal signalling for lateral bud initiation.**

**a**, Tissue of the budding zone expressing green fluorescent protein (GFP) was treated with purified recombinant *H. magnipapillata* Ndr protein and transplanted into the identical region of untreated wild-type *Hydra*; BSA was used as a control. **b**, Quantification of the effect of Ndr treatment on bud formation. **c**, Enhancement of bud-forming capacity by Ndr was confirmed by overexpressing Ndr after electroporation of cells with Ndr-expressing plasmid. **d**, Experimental scheme showing Alk4/5/7 inhibitor treatment and siRNA-mediated Ndr or Pitx knockdown. Note that Ndr and Pitx, but not ActL1, are expressed (blue) at stage 1. The lines in orange and blue show Ndr inhibition starting at bud stages 1 and 3, respectively. **e**, Treatment with the inhibitors A-83-01 (50 nM), SB-505124 (10  $\mu$ M), LY-364947 (100 nM) or SD-208 (2.5  $\mu$ M) inhibits bud induction when treatment started at bud stage 1. DMSO, dimethylsulphoxide. **f**, A-83-01 inhibited bud induction (B1 and B2 buds). **g**, A-83-01 had no effect on budding when treatment started after bud stage 3. **h**, **i**, Knockdown of Ndr (**h**) or Pitx (**i**) using siRNA at bud stage 1 suppressed bud induction. The kinetics of bud inhibition in siRNA-transfected animals are shown. **b**, **c**, **e**, **f**–**i**, The data are presented as mean  $\pm$  s.e.m. \*,  $P < 0.05$  compared with control (BSA, DMSO or GFP siRNA); one-sided *t*-test.

*Bmp2/4*, *Bmp5*–*8* (ref. 8) and Activin-related genes were expressed ubiquitously or mainly at the oral side of polyps and did not show any sign of lateral asymmetry within the budding zone (Extended Data Fig. 7). In bilaterians, Nodal and its target transcription factor Pitx2 constitute a co-expressed signalling unit<sup>9,10</sup>. The *H. magnipapillata* Pitx gene was expressed unilaterally in the earliest bud stage, similarly to Ndr (Fig. 1e and Extended Data Fig. 7). The other Nodal target genes in *H. magnipapillata*, *FoxA* and *Lhx1/5* (ref. 11), were also expressed in the developing bud but not in the presumptive budding zone (Extended Data Fig. 7).



**Figure 3 | Autoregulatory feedback loop of Nodal signalling and asymmetrical patterning of the budding zone.** **a**, Reversible suppression of *H. magnipapillata* Ndr, Pitx and Bra1 expression (blue) by A-83-01 treatment. Ndr, Pitx and Bra1 expression is suppressed by A-83-01 but strongly upregulated 24 h after inhibitor removal. The numbers on the images denote the number of Ndr-, Pitx- or Bra1-expressing polyps/the total number of polyps. **b**, Expanded bud formation after inhibitor removal. Note the perturbation of the biradial asymmetry of the budding zone after A-83-01 removal, with an expansion of Pitx expression (blue) and multiple signalling centres expressing the head organizer genes Wnt3a and Bra1 (both blue) in a spot-like pattern. The numbers denote the number of biradial, expanded or radial polyps/the total number of polyps.

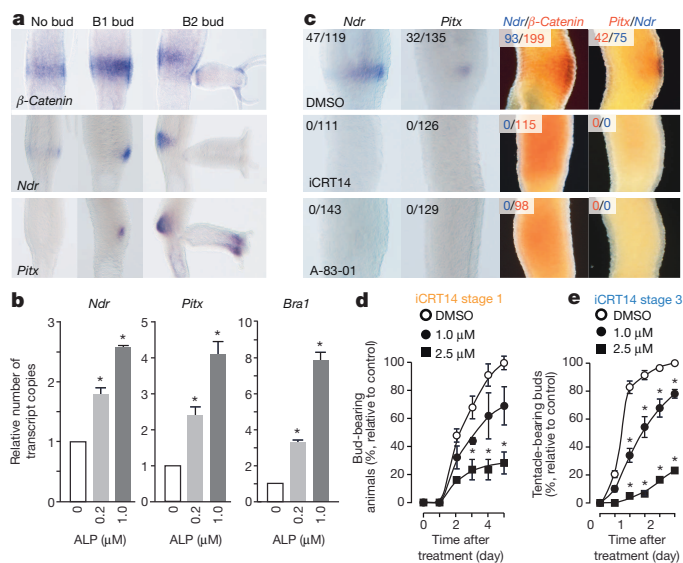
To investigate the function of *H. magnipapillata* Ndr signalling, we examined the effect of recombinant Ndr on bud formation. When tissue explants of the budding zone from budless animals were treated with purified recombinant Ndr protein and transplanted into the corresponding region of untreated hosts, the rate of budding was markedly higher than for BSA-treated control tissue (Fig. 2a, b). We also observed an activation of bud formation in animals in which Ndr was overexpressed after electroporation of an Ndr-expressing plasmid (Fig. 2c). While Ndr-transfected cells were observed throughout the body axis, the overexpression increased the budding rate only from the budding zone and did not induce ectopic bud formation (Extended Data Fig. 8). This finding suggests that only the budding zone contains cells competent for Ndr activity and/or that additional signalling pathways are required for bud induction.

The requirement for Ndr for bud formation was also examined by using chemical inhibitors of Activin-receptor-like kinase 4 (Alk4)/5/7. Nodal, Activin and Tgf- $\beta$  share Alk4/5/7 as receptors, and their activities can be blocked through specific interference by administering A-83-01, SB-505124, LY-364947 and SD-208 (ref. 12). While none of the *Hydra* Activin-related genes was expressed before the bud became visible at stage 3 (Fig. 2d and Extended Data Fig. 7b), *H. magnipapillata* Ndr and Pitx were exclusively expressed in the bud anlage at stage 1–2 (Fig. 2d). When we treated polyps with Alk4/5/7 inhibitors from bud stage 1–2 onwards (Fig. 2d, orange line), the bud formation was significantly inhibited, with the strongest effect with A-83-01 administration (Fig. 2e, f and Extended Data Fig. 9a–d). By comparison, inhibitor treatment from stage 3 onwards yielded normal bud formation (Fig. 2g). The crucial function of Ndr and Pitx in budding was confirmed by short interfering RNA (siRNA)-mediated gene knockdown experiments (Fig. 2h, i, and Extended Data Fig. 10). These data show that the function of Ndr and Pitx is essential for bud initiation.

In animals that were treated with the inhibitor A-83-01, we found complete downregulation of the expression of *H. magnipapillata* Ndr, Pitx and *Brachyury1* (*Bra1*), an early head marker, at the bud anlage (Fig. 3a). This indicates that Nodal signalling is required for the asymmetrical expression of Pitx and Bra1 and suggests an autoregulatory control of Ndr expression. Removal of the inhibitor resulted in an unexpectedly strong response. More than half of the treated polyps did not regain their normal, biradial, budding pattern. Instead, they showed

expanded expression domains of *Ndr*, *Pitx* and *Bra1* around the budding zone (Fig. 3a). The resultant bud tissue exhibited expanded and/or multiple expression domains of *Wnt3a*, *Bra1* and *Pitx* (Fig. 3b), as well as an increase in the number of head-specific tentacles (Supplementary Table 2). These data not only show that *Ndr*, *Pitx* and *Bra1* are downstream targets of Nodal signalling in *Hydra* but also that the localized activation of *Ndr* signalling at one side of the ring-like budding zone is crucial for inducing a bud anlage. The expanded expression of *Ndr* and its target genes suggests that, in addition to the autoregulatory feedback loop of Nodal signalling, an inhibitory factor is active. The expression of this putative inhibitor restricts Nodal activity to the bud anlage. This hypothesis is reminiscent of the Nodal–Lefty feedback interaction, which belongs to the best-known activator–inhibitor systems in vertebrates<sup>13,14</sup>. Lefty has so far been found only in deuterostomes (Extended Data Figs 1 and 2). Despite the existence of various *Cerberus*-related sequences in *Hydra* and other cnidarians (Extended Data Figs 1 and 4a), whole mount *in situ* hybridization (WISH) data exclude the involvement of these genes in the asymmetrical induction of the bud anlage (Extended Data Fig. 7b). Thus, the molecular nature of Nodal antagonism in *Hydra* remains to be identified.

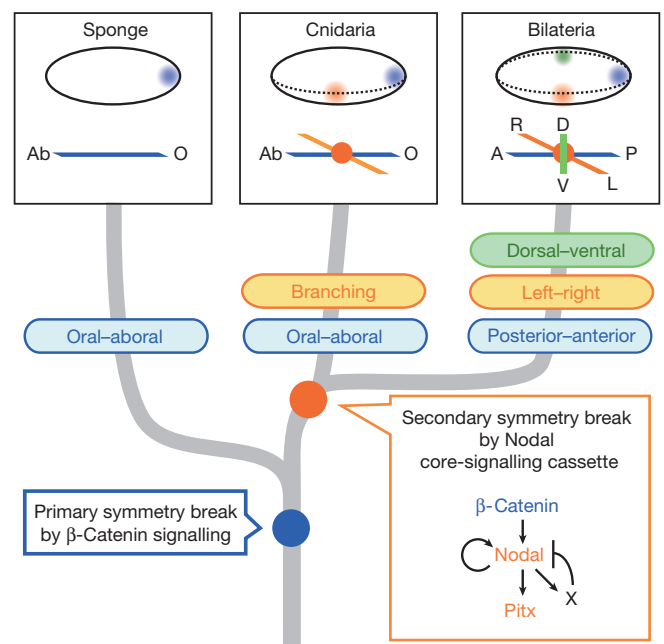
Next, we addressed the question of regulatory genes that act upstream of Nodal signalling. One candidate is  $\beta$ -Catenin, which is upregulated in a ring-like expression domain that defines the budding zone<sup>15</sup>. It has been postulated that within this ring, a single peak of activation is created by lateral inhibition, which induces the evagination of the bud<sup>15</sup>. Figure 4a shows that *H. magnipapillata* *Ndr* exhibits a similar expression pattern to  $\beta$ -Catenin in each prospective budding zone but



**Figure 4 |  $\beta$ -Catenin signalling is essential for *H. magnipapillata* *Ndr* expression and bud formation.** **a**, Co-expression of *H. magnipapillata*  $\beta$ -Catenin, *Ndr* and *Pitx* in the presumptive budding zone and early bud (B1 and B2 buds, stage 1). At bud stage 1, the expression of  $\beta$ -Catenin and *Ndr* become restricted towards the bud anlage, resulting in localized *Pitx* expression. **b**, Increased expression of *Ndr*, *Pitx* and *Bra1* in polyps treated with the GSK3 inhibitor ALP for 24 h. **c**, Left, Inhibition of *Ndr* and *Pitx* expression (both blue) by treatment with the  $\beta$ -Catenin inhibitor iCRT14 or the Alk4/5/7 inhibitor A-83-01. The numbers on the images denote the number of positive animals/the total number of animals. Right, Double WISH staining of *Ndr* (blue) and  $\beta$ -Catenin (red) or *Pitx* (red) expression after treatment with iCRT14 or A-83-01. The numbers denote the number of *Ndr*-positive polyps/the number of  $\beta$ -Catenin-positive polyps or the number of *Pitx*-positive polyps/the number of *Ndr*-positive polyps. **d**, **e**, Dose-dependent inhibition of bud initiation (**d**) and bud development (**e**) by iCRT14 treatment at stages 1 and 3, respectively. **b**, **d**, **e**, The data are presented as mean  $\pm$  s.e.m.; \*,  $P < 0.05$  compared with control (0  $\mu$ M ALP or DMSO); one-sided *t*-test, (for further details, see Methods).

then becomes restricted to the side of the emerging bud (stage 1), where *Pitx* becomes transcriptionally activated. This finding suggests that *Ndr* and *Pitx* act downstream of  $\beta$ -Catenin signalling in initiating the site of bud evagination. We tested this hypothesis by ectopically activating Wnt– $\beta$ -Catenin signalling with the GSK3 inhibitor alsterpaullone (ALP)<sup>16</sup>. In ALP-treated polyps, *Ndr*, *Pitx* and *Bra1* expression were upregulated in a dose-dependent manner (Fig. 4b). Inhibition of  $\beta$ -Catenin by treating animals with the  $\beta$ -Catenin inhibitor iCRT14 blocked the expression of these target genes in the budding region (Fig. 4c). Double WISH analyses demonstrated that the  $\beta$ -Catenin, *Ndr* and *Pitx* have an overlapping expression pattern at the bud anlage (Fig. 4c). Inhibitors of  $\beta$ -Catenin (iCRT14) and Nodal (A-83-01) signalling abrogated *Ndr* and *Pitx* expression in the budding zone (Fig. 4c). Quantitative PCR (qPCR) data show that this inhibitory effect is concentration dependent for *Pitx* (Extended Data Fig. 9e). This finding was confirmed by the concentration-dependent inhibition of bud formation by treating animals with iCRT14 at bud stage 1 (Fig. 4d). Inhibition of  $\beta$ -Catenin signalling at bud stages 3 and later also interfered with bud development (Fig. 4e). This finding substantiates previous reports suggesting a function of  $\beta$ -Catenin in the *Hydra* organizer<sup>15,16</sup>. It also indicates that Nodal signalling functions downstream of Wnt– $\beta$ -Catenin signalling only during bud initiation to induce the head organizer in an asymmetrical and biradial pattern.

In summary, our data show that  $\beta$ -Catenin and Nodal–*Pitx* signalling act together in a core-signalling cassette that has the potential to break symmetry in a morphogenetic field characterized by high  $\beta$ -Catenin activity (Fig. 5). Downstream of Wnt– $\beta$ -Catenin signalling, the Nodal–*Pitx* pathway induces biradial budding asymmetry of *Hydra*. The same pathway acts in left–right axis formation in vertebrates and snails<sup>10,17–20</sup>, in which the symmetry-breaking event is upstream of



**Figure 5 | Nodal signalling and secondary body axes of eumetazoans.** The primary Wnt– $\beta$ -Catenin axis (blue) was established by a primary symmetry break in the last common metazoan ancestor and led to the anterior (aboral, Ab)–posterior (oral, O) axis pattern. In the common ancestor of cnidarians and bilaterians, the core-signalling cassette comprising  $\beta$ -Catenin–Nodal–*Pitx*-signalling components evolved and induced a secondary symmetry break and further body axes (orange). In the branching cnidarian *Hydra*, the signalling cassette was employed to specify an orthogonal axis inducing lateral budding. In modern bilaterians, this signalling cassette was deployed to specify the left–right (L–R) axis. This signalling cassette also has a crucial role, in a concerted action on Bmp signalling, for dorsal–ventral (D–V) axis patterning, by inducing Bmp antagonists at the dorsal organizer in vertebrates. A, anterior; P, posterior.



Nodal expression and a loss of Nodal signalling causes randomization of asymmetry<sup>3,10</sup>. Likewise, we found a randomization of asymmetry when Nodal was overexpressed after inhibitor removal and thus overrode the given asymmetry cues (Fig. 3). Our results also provide new insight into the evolution of animal body plans. So far, Nodal signalling has not been considered to have a prominent function in the early evolution of metazoan body axes. We propose that the  $\beta$ -Catenin–Nodal–Pitx core-signalling cassette evolved in early metazoans before the cnidarian–bilaterian split (Fig. 5). Accordingly, the last common ancestor of cnidarians and bilaterians exhibited a major oral–aboral axis patterned by Wnt– $\beta$ -Catenin signalling<sup>21</sup> and was able to break radial symmetry by using Nodal signalling. Nodal signalling might then have been differentially co-opted in animal evolution. In *Hydra* and other cnidarians, Nodal signalling causes biradial branching and colony formation. In bilaterians, Nodal signalling has a basic role in establishing left–right symmetry (see above), is involved in the primary axial patterning of sea urchins and amphioxus<sup>17,18,22</sup> and acts in the formation of the organizer in chordates<sup>23,24</sup>. Further studies are needed to identify how Nodal has contributed to the evolutionarily more complex morphogenetic program for the organizer and for axis formation in bilaterians.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 26 May 2013; accepted 7 July 2014.

Published online 24 August 2014.

- Niehrs, C. On growth and form: a Cartesian coordinate system of Wnt and BMP signaling specifies bilaterian body axes. *Development* **137**, 845–857 (2010).
- Shen, M. M. Nodal signaling: developmental roles and regulation. *Development* **134**, 1023–1034 (2007).
- Schier, A. F. Nodal morphogens. *Cold Spring Harb. Perspect. Biol.* **1**, a003459 (2009).
- Meinhardt, H. Primary body axes of vertebrates: generation of a near-Cartesian coordinate system and the role of Spemann-type organizer. *Dev. Dyn.* **235**, 2907–2919 (2006).
- Brusca, R. C. & Brusca, G. J. *Invertebrates* 2nd edn (Sinauer, 2002).
- Nielsen, C. *Animal Evolution: Interrelationships of the Living Phyla* 3rd edn (Oxford Univ. Press, 2012).
- Baird, R. V. & Burnett, A. L. Observations on the discovery of a dorso-ventral axis in *Hydra*. *J. Embryol. Exp. Morphol.* **17**, 35–81 (1967).
- Reinhardt, B., Broun, M., Blitz, I. L. & Bode, H. R. HyBMP5–8b, a BMP5–8 orthologue, acts during axial patterning and tentacle formation in *hydra*. *Dev. Biol.* **267**, 43–59 (2004).
- Chea, H. K., Wright, C. V. & Swalla, B. J. Nodal signaling and the evolution of deuterostome gastrulation. *Dev. Dyn.* **234**, 269–278 (2005).
- Grande, C. & Patel, N. H. Nodal signalling is involved in left–right asymmetry in snails. *Nature* **457**, 1007–1011 (2009).
- Srivastava, M. *et al.* Early evolution of the LIM homeobox gene family. *BMC Biol.* **8**, 4 (2010).
- Vogt, J., Traynor, R. & Sapkota, G. P. The specificities of small molecule inhibitors of the TGF $\beta$  and BMP pathways. *Cell. Signal.* **23**, 1831–1842 (2011).
- Meinhardt, H. Organizer and axes formation as a self-organizing process. *Int. J. Dev. Biol.* **45**, 177–188 (2001).
- Hamada, H., Meno, C., Watanabe, D. & Saijoh, Y. Establishment of vertebrate left–right asymmetry. *Nature Rev. Genet.* **3**, 103–113 (2002).
- Hobmayer, B. *et al.* WNT signalling molecules act in axis formation in the diploblastic metazoan *Hydra*. *Nature* **407**, 186–189 (2000).
- Broun, M., Gee, L., Reinhardt, B. & Bode, H. R. Formation of the head organizer in *hydra* involves the canonical Wnt pathway. *Development* **132**, 2907–2916 (2005).
- Duboc, V., Röttinger, E., Lapraz, F., Besnardeau, L. & Lepage, T. Left–right asymmetry in the sea urchin embryo is regulated by Nodal signaling on the right side. *Dev. Cell* **9**, 147–158 (2005).
- Yu, J. K., Holland, L. Z. & Holland, N. D. An amphioxus *nodal* gene (*AmphiNodal*) with early symmetrical expression in the organizer and mesoderm and later asymmetrical expression associated with left–right axis formation. *Evol. Dev.* **4**, 418–425 (2002).
- Shiratori, H. & Hamada, H. The left–right axis in the mouse: from origin to morphology. *Development* **133**, 2095–2104 (2006).
- Yamamoto, M. *et al.* Nodal antagonists regulate formation of the anteroposterior axis of the mouse embryo. *Nature* **428**, 387–392 (2004).
- Holstein, T. W. The evolution of the Wnt pathway. *Cold Spring Harb. Perspect. Biol.* **4**, a007922 (2012).
- Warner, J. F., Lyons, D. C. & McClay, D. R. Left–right asymmetry in the sea urchin embryo: BMP and the asymmetrical origins of the adult. *PLoS Biol.* **10**, e1001404 (2012).
- Inui, M. *et al.* Self-regulation of the head-inducing properties of the Spemann organizer. *Proc. Natl Acad. Sci. USA* **109**, 15354–15359 (2012).
- Feldman, B. *et al.* Zebrafish organizer development and germ-layer formation require nodal-related signals. *Nature* **395**, 181–185 (1998).
- Kühn, A. Entwicklungsgeschichte und Verwandtschaftsbeziehungen der Hydrozoen: Die Hydroiden. *Ergebnisse und Fortschritte der Zoologie* **4**, 1–284 (1914).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** We thank R. Steele for providing the transgenic *Hydra* strains before publication, for setting up the *Cladonema pacificum* genome Blast server and for permission to provide the URL; R. Yamada, T. Katsuki and R. Greenspan for providing unpublished *C. pacificum* sequence data; B. Hobmayer for providing a *Hydra* scheme from an unpublished manuscript; and H. Yuzawa-Watanabe for technical assistance. We thank O. Simakov, H. Meinhardt, I. Somorjai and B. Hobmayer for discussions and comments on the manuscript. This study was supported by the TOYBO Biotechnology Foundation and the Alexander von Humboldt Foundation (initial fellowships to H.W.), the Heidelberg Excellence Cluster Cellular Networks, and grants from the German science foundation (DFG) to T.W.H. (FOR 1036/TP1 and SFB 873/A1) and S.O. (FOR 1036/TP2).

**Author Contributions** H.W. and T.W.H. designed the research. H.W. performed most of the experiments. A.K., Y.K. and N.L.-L. performed the WISH experiments. S.O. performed the protein biochemistry experiments. S.K.H. performed the qPCR. H.A.S. and T.W.H. performed the phylogenomic and sequence analyses. H.W., S.O. and T.W.H. wrote the manuscript.

**Author Information** Multiple sequence alignments and phylogenetic trees for the maximum likelihood phylogenies of the metazoan *Tgf- $\beta$* , *Cerberus*, *Dan* and *Gremelin*, *Lhx* and *Lim*, *Smad*, and *Prd* and *Pitx* gene families have been deposited with TreeBASE (<http://treebase.org>) under the Study ID S16190. Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to T.W.H. ([thomas.holstein@cos.uni-heidelberg.de](mailto:thomas.holstein@cos.uni-heidelberg.de)) or H.W. ([hiroshi.watanabe@cos.uni-heidelberg.de](mailto:hiroshi.watanabe@cos.uni-heidelberg.de)).

## METHODS

**Molecular phylogenetic analysis.** Sequences were aligned using MAFFT multiple sequence alignment and were checked manually. Phylogenetic trees were reconstructed using BioNJ and maximum likelihood as implemented in IQPNNI v3.3 (ref. 26) (using 600 iterations and the WAG model<sup>27</sup>) and also using PhyML v3.0 or v3.1 (ref. 28) (using the LG model<sup>29</sup>). Bootstrap support values were obtained from 100 bootstrap samples (if not stated otherwise) with the same parameters (using the bootstrap option -bs in IQPNNI v3.3) and were summarized using a relative majority consensus as implemented in TREE-PUZZLE v5.3 (ref. 30). In all cases, rate heterogeneity was modelled using four discrete gamma rate categories. The NCBI accession numbers for all genes annotated in this study are AB823860 to AB824011. *C. pacificum* sequences are accessible at <http://polyp.biochem.uci.edu/blast/>.

**Culture of *Hydra* strains.** Experiments were carried out with *H. magnipapillata* (strain 105) or *Hydra vulgaris* (strain AEP), which were cultured as described previously<sup>31</sup>. To obtain *Hydra* polyps bearing the bud anlage, we collected living animals using the following procedure. Newly detached small polyps (B bud generation, Fig. 1c) were collected within a defined time window (that is, every 6 h). These animals were cultured at a density of 1–1.3 hydra cm<sup>-2</sup> and fed and washed daily. Under these conditions, the detached daughter polyps start the formation of new buds (C bud generation, Fig. 1c) after 70–76 h. The budless *Hydra* polyps were treated with inhibitors 24–48 h after detachment and were fixed for WISH analyses at 72 h after detachment.

**Gene isolation.** Partial open reading frame (ORF) sequences of *Hydra* genes, including those encoding Tgf- $\beta$  and Dan family proteins and transcription factors, were obtained by database searches on GenBank and Hydrazome. Gene-specific primers were designed to recover both 3' and 5' RACE (rapid amplification of cDNA ends) fragments using the GeneRacer kit (Invitrogen) with annealing temperatures between 60 °C and 65 °C. The RACE products were cloned using pGEM-T (Promega) and sequenced. Overlapping 5' and 3' RACE fragments were aligned to obtain messenger RNA sequences.

**WISH and qPCR.** WISH studies were performed as described previously<sup>31,32</sup>. For qPCR experiments, total RNA was isolated from embryos with the use of an RNeasy kit (QIAGEN) and reverse transcribed into cDNA. qPCR analysis was carried out using a DNA Engine Thermal Cycler equipped with a Chromo4 Real-Time Detector (Bio-Rad). The qPCR primer sequences used to amplify *H. magnipapillata* cDNA are as follows: *Ndr* forward, 5'-GCACATCCGCTGACAGTTAC-3'; *Ndr* reverse, ACTCTGTTGGAACACAGCAAGG-3'; *Pitx* forward, 5'-CGAGCAGCCAAGC TTCTAATGT-3'; *Pitx* reverse, 5'-CAGCTGTGTATCCGCTCGTA-3'; *Bra1* forward, 5'-TCACAGGTGGAATACGTAAATGGA-3'; *Bra1* reverse, 5'-ATCGG GCTTTTCATCCAATGTGT-3'.

**Inhibitor treatment.** To examine the effect of Nodal and Activin inhibitors on bud induction, young budless animals (1–2 days after detachment) were treated with an increasing concentration of A-83-01 (Calbiochem), SB-505124 (Sigma), LY-364947 (Sigma) or SD-208 (Sigma). The data shown in Fig. 2e and Extended Data Fig. 9 are presented as the mean  $\pm$  s.e.m. of four experiments ( $n = 82$  dimethylsulphoxide (DMSO),  $n = 96$  A-83-01,  $n = 83$  SB-505124,  $n = 87$  LY-364947 and  $n = 81$  SD-208). In the experiment shown in Fig. 2f, 120 and 119 budless animals were used in the control (DMSO) and experimental (40 nM A-83-01) condition, respectively. The data are presented as the mean  $\pm$  s.e.m. of four experiments. To check the effect of A-83-01 on bud development, animals bearing buds at stage 3 were treated with 40 nM A-83-01 ( $n = 71$  DMSO and  $n = 66$  A-83-01 in three independent experiments). To induce expanded and radialized bud formation, budless animals were treated with 40 nM A-83-01 for 3 days and cultured for 1 day (Fig. 3a) or 3 days (Fig. 3b) without the inhibitor. The number of buds showing biradial, expanded or radial morphologies ( $n = 222$  control and  $n = 324$  A-83-01 in three independent experiments) and the number of tentacles per polyp ( $n = 60$  control and  $n = 28$  A-83-01 in two independent experiments) were scored. Animals were fed and washed each day in hydra medium containing the inhibitors. The GSK3 inhibitor alsterpaullone (ALP) (Calbiochem) was used at 0.2  $\mu$ M and 1  $\mu$ M. For the qPCR experiments shown in Fig. 4b, budless animals were treated with ALP for 24 h ( $n = 4$ ). Animals were treated with 1–2.5  $\mu$ M and 5  $\mu$ M  $\beta$ -Catenin inhibitor iCRT14 (Sigma) to examine the effect on budding and *Ndr* and *Pitx* expression, respectively. Dose-dependent iCRT14 inhibition of bud initiation ( $n = 48$  DMSO,  $n = 47$  1  $\mu$ M iCRT14 and  $n = 48$  2.5  $\mu$ M iCRT14) and bud development ( $n = 58$  DMSO,  $n = 59$  1  $\mu$ M iCRT14 and  $n = 60$  2.5  $\mu$ M iCRT14) were examined in three independent experiments for each. In all experiments using chemical inhibitors, the control medium contained an equal concentration (less than 0.1% (v/v)) of DMSO.

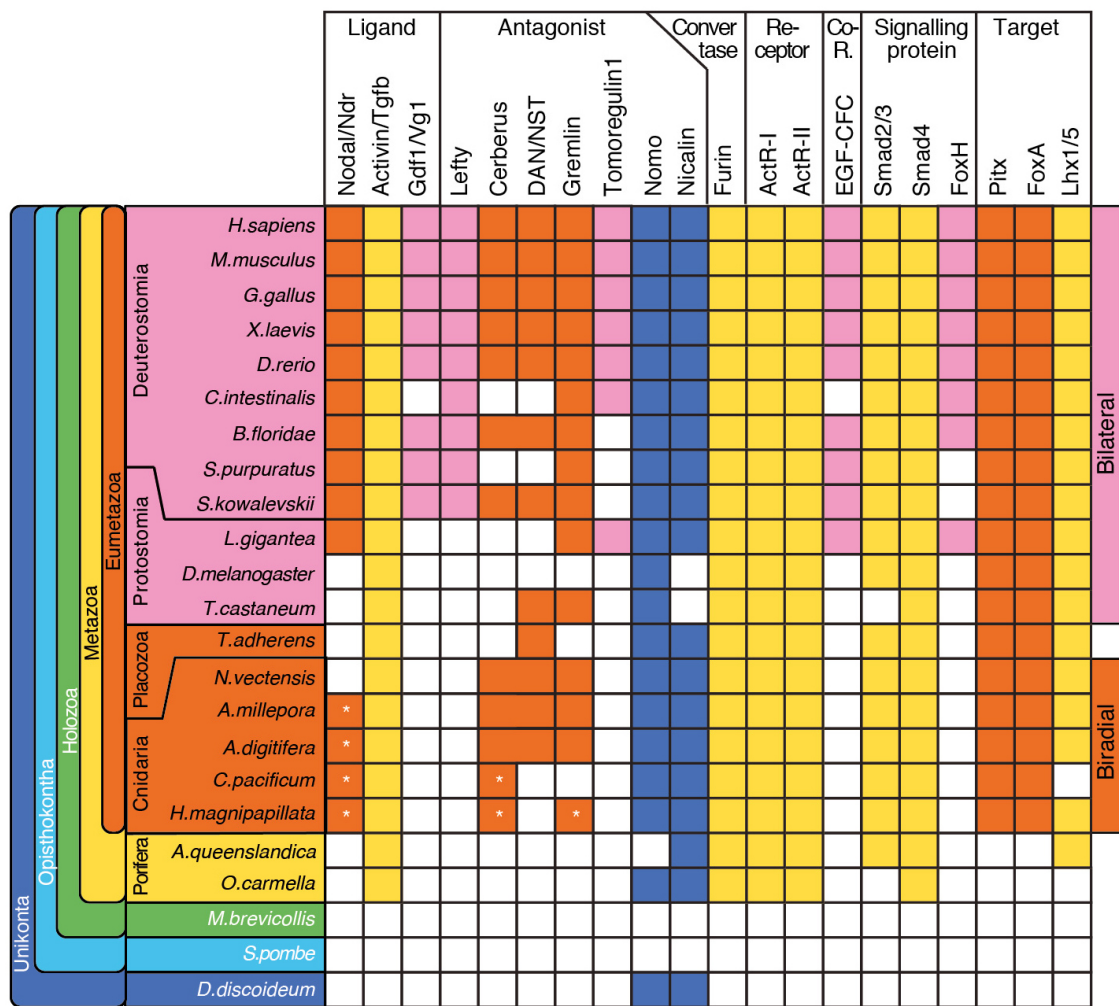
**Treatment with recombinant *H. magnipapillata* *Ndr* protein.** Recombinant His-tagged *Ndr* protein was expressed in *Escherichia coli* and purified, and the effect on bud formation was examined as described previously with minor modifications<sup>31</sup>. For *Ndr* experiments, a piece of mid-gastric tissue (one eighth of the body column) was isolated from budless animals of transgenic *H. vulgaris* AEP in which green

fluorescent protein (GFP) was expressed in ectodermal epithelial cells under the control of the *Hydra Actin* promoter and treated with BSA (control) or recombinant *Ndr* protein for 3 h in hydra medium containing 20% DMEM (Sigma). After treatment, the tissue explants were transplanted into untreated budless polyps of the normal AEP strain ( $n = 58$  for BSA and  $n = 58$  for *Ndr* in four independent experiments).

**Overexpression and knockdown of *H. magnipapillata* *Ndr* and *Pitx*.** For the expression plasmid, the ORF of *Ndr* was cloned into the vector pBSAA under the control of the *Hydra Actin* promoter<sup>33</sup>. For the control experiment, empty pBSAA or pBSAA-GFP vectors were used. The numbers of animals used was 432 for the controls and 434 for the *Ndr* studies in 12 independent experiments. The siRNAs specific for *GFP* (AAGATGGAACATTCTTGGAC), *Ndr* (*Ndr* siRNA2, AAGG AACAATATTCTACGA; and *Ndr* siRNA3, TGGATTACATTGGAAGTAAC A) and *Pitx* (*Pitx* siRNA1, ATGTGGAAAATAAGAGATACG; and *Pitx* siRNA2, GAGTTGATAATCGCATAAGAAGA) were purchased from QIAGEN. Transfection of the plasmid or siRNAs was carried out as reported previously<sup>34</sup> with minor modifications. As *Ndr* and *Pitx* are expressed in different germ layers (that is, in the ectoderm and endoderm, respectively), we used transgenic *Hydra* strains in which either the ectoderm or endoderm was labelled with GFP or red fluorescent protein (RFP), respectively<sup>35</sup>. Budless animals were collected into a Petri dish from cultures fed every second day. The animals were washed five times with Milli-Q water and left for 30–60 min. Polyps (20–25) were placed into an electroporation cuvette with a 4-mm gap (Bio-Rad), and as much water was removed as possible. After adding 200  $\mu$ l sterilized 10 mM HEPES (pH 7.0) containing 10  $\mu$ g expression vector or 2  $\mu$ M siRNA, the cuvettes were shaken by tapping ten times to evenly distribute the animals and the plasmid or siRNA. The cuvettes were left for 5 min to allow the animals to relax and extend. A Bio-Rad Gene Pulser II electroporation system equipped with a radio-frequency (RF) module was adjusted (for plasmids, 60 V, 1 burst, 50 ms burst duration; for siRNA, 50 V, 1 burst, 10 ms burst duration). Immediately after administering each pulse, 500  $\mu$ l restoration medium containing 80% hydra medium and 20% hyperosmotic dissociation medium (6 mM CaCl<sub>2</sub>, 1.2 mM MgSO<sub>4</sub>, 3.6 mM KCl, 12.5 mM N-Tris-[hydroxymethyl]methyl-2-aminoethanesulphonic acid, 6 mM sodium pyruvate, 6 mM sodium citrate, 6 mM glucose, and 50 mg ml<sup>-1</sup> rifampicin, pH 6.9) was added to the cuvette. Animals were then carefully transferred to Petri dishes (6-cm diameter) containing 10 ml restoration medium and incubated for 1 day to allow recovery. Viable polyps were separated from cell debris and transferred to new Petri dishes containing 100% hydra medium, cultured for 1 day and used for experiments. In the siRNA-mediated *Ndr* knockdown experiments, we performed six independent experiments ( $n = 139$  for *GFP*-specific siRNA,  $n = 138$  for *Ndr* siRNA2 and  $n = 136$  for *Ndr* siRNA3). For *Pitx* knockdown, we performed three independent experiments ( $n = 36$  for *GFP*-specific siRNA,  $n = 40$  for *Pitx* siRNA1 and  $n = 35$  for *Pitx* siRNA2).

**Statistics.** The precise numbers of experimental animals are reported above. All statistical analyses were performed with a one-sided, unpaired Student's *t*-test. The data are shown as the mean  $\pm$  s.e.m. No animal or sample was excluded from the analysis. Sample randomization and blinding were not applied, as the treatments used were not targeted at specific sites or organs of the animal.  $P < 0.05$  was considered to indicate statistical significance.

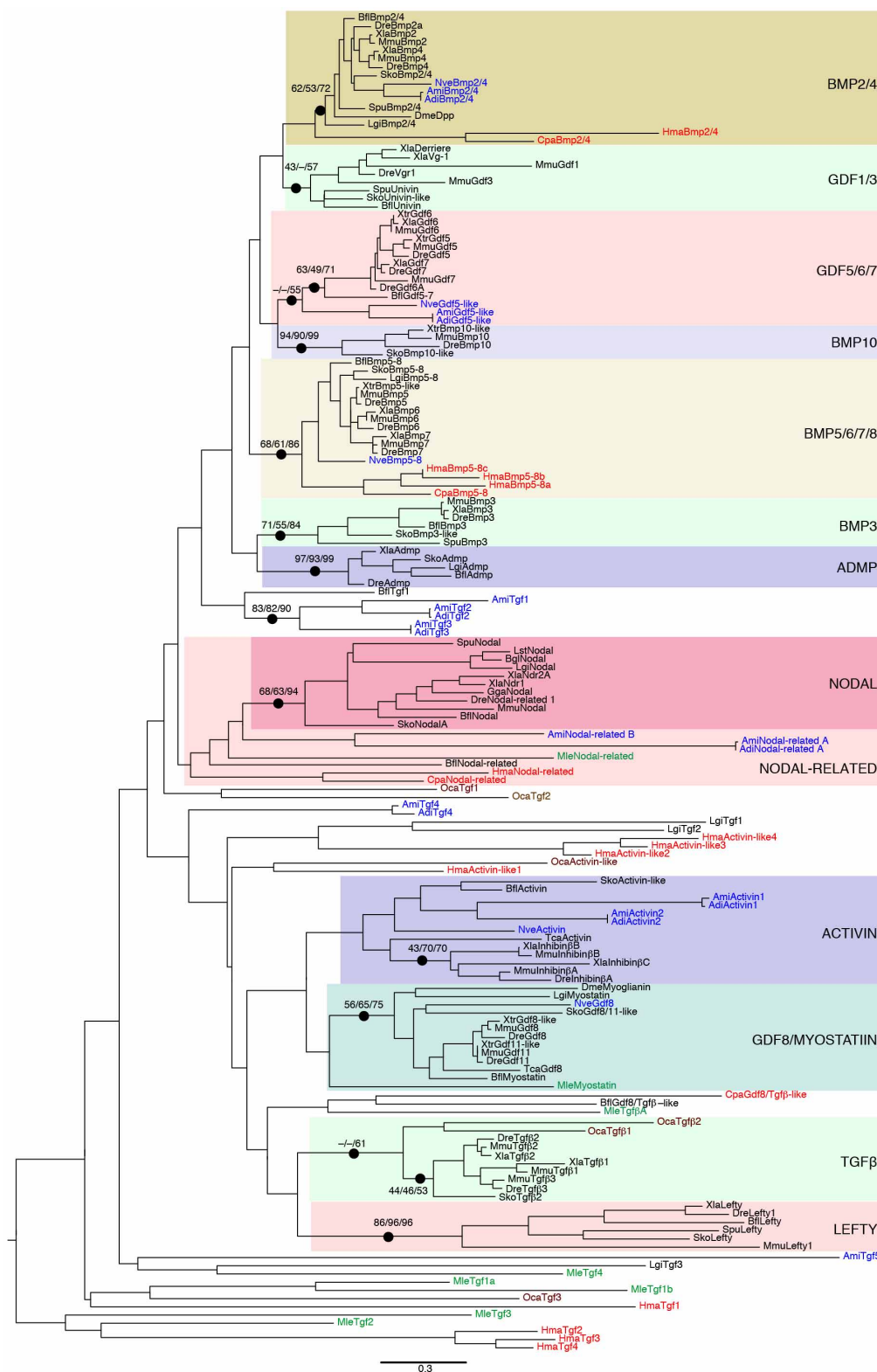
- Minh, B. Q., Vinh, L. S., von Haeseler, A. & Schmidt, H. A. pIQPNNI: parallel reconstruction of large maximum likelihood phylogenies. *Bioinformatics* **21**, 3794–3796 (2005).
- Whelan, S. & Goldman, N. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.* **18**, 691–699 (2001).
- Guindon, S. et al. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).
- Le, S. Q. & Gascuel, O. An improved general amino acid replacement matrix. *Mol. Biol. Evol.* **25**, 1307–1320 (2008).
- Schmidt, H. A., Strimmer, K., Vingron, M. & von Haeseler, A. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* **18**, 502–504 (2002).
- Lengfeld, T. et al. Multiple Wnts are involved in *Hydra* organizer formation and regeneration. *Dev. Biol.* **330**, 186–199 (2009).
- Bode, H., Lengfeld, T., Hobmayer, B. & Holstein, T. W. Detection of expression patterns in *Hydra* pattern formation. *Methods Mol. Biol.* **469**, 69–84 (2008).
- Nakamura, Y., Tsiarlis, C. D., Özbek, S. & Holstein, T. W. Autoregulatory and repressive inputs localize *Hydra* Wnt3 to the head organizer. *Proc. Natl Acad. Sci. USA* **108**, 9137–9142 (2011).
- Khalturin, K. et al. A novel gene family controls species-specific morphological traits in *Hydra*. *PLoS Biol.* **6**, e278 (2008).
- Glauber, K. M. et al. A small molecule screen identifies a novel compound that induces a homeotic transformation in *Hydra*. *Development* **140**, 4788–4796 (2013).



**Extended Data Figure 1 | Phylogenomic distribution of Nodal-signalling genes.** Rows, species. Columns, gene members involved in Nodal signalling; white asterisks indicate low statistical support in phylogenomic analyses

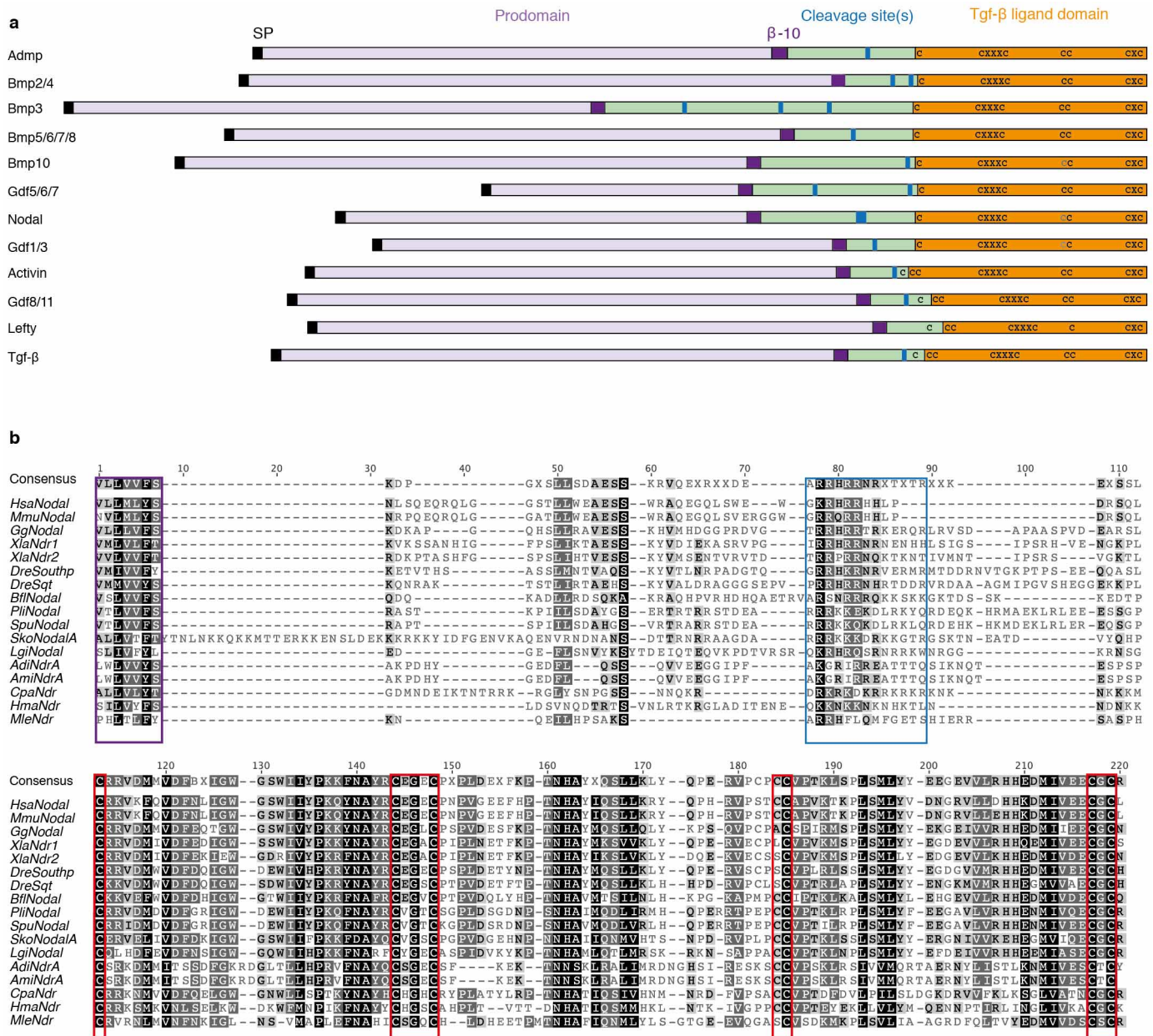
(Extended Data Figs 2, 4–6). Genes labelled with pink evolved in Bilateria; orange in Eumetazoa; yellow, in Metazoa; blue, in Unikonta.

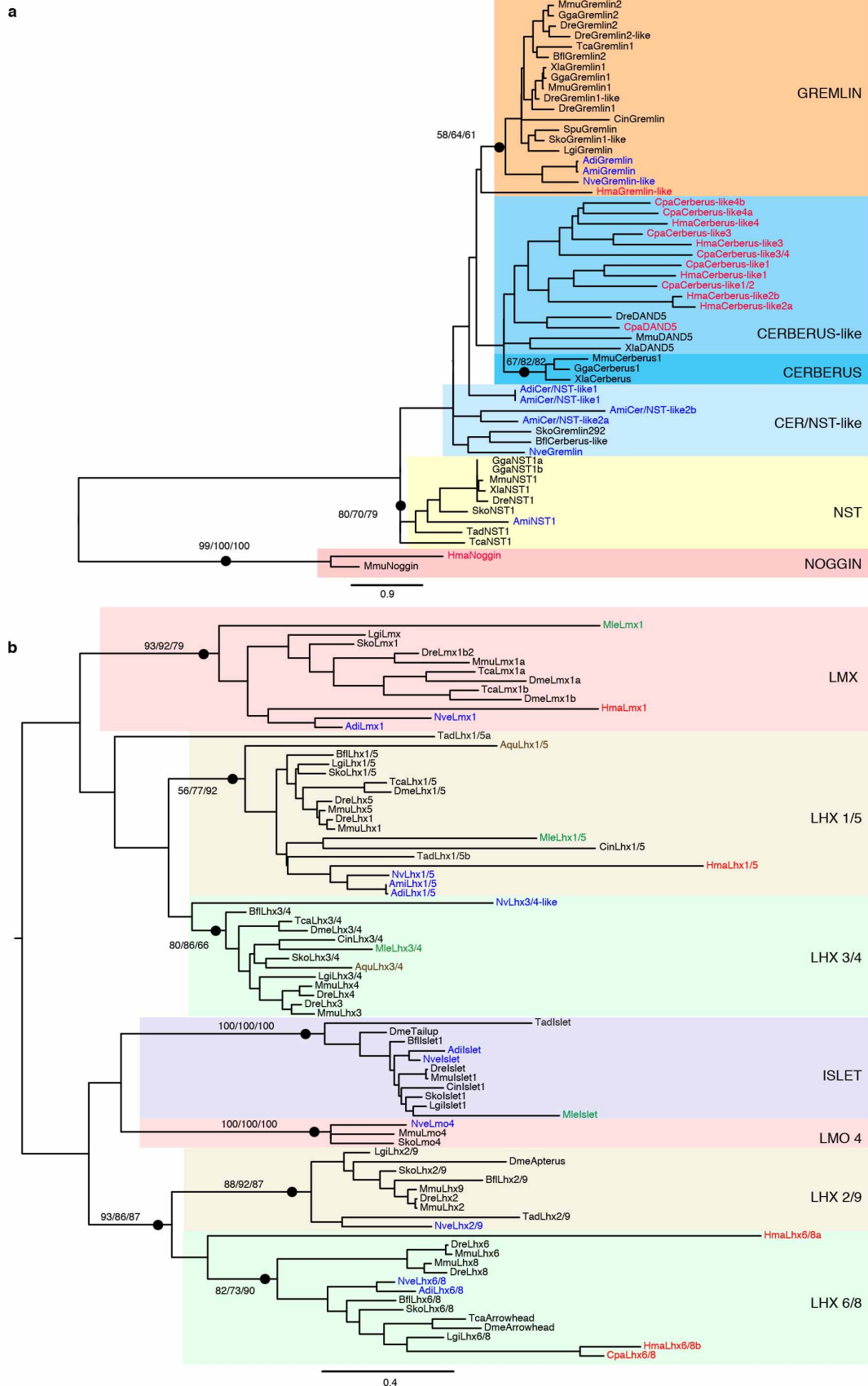




**Extended Data Figure 2 | Maximum likelihood phylogeny of the metazoan Tgf-β gene superfamily.** *H. magnipapillata* (*Hma*) and *Cladonema pacificum* (*Cpa*) genes are shown in red, and *Nematostella vectensis* (*Nve*), *Acropora digitifera* (*Adi*) and *Acropora millepora* (*Ami*) genes are shown in blue. The numbers at the branches indicate the bootstrap supports from 200 BioNJ, 200 IQPNNI and 100 PhyML bootstrap trees, respectively. There was reasonable support for the cnidarian members of the *Bmp2/4*, *Bmp5–8*, *Gdf5–7* and *Gdf8* (Myostatin) gene families. There are no cnidarian members of the *Bmp3*, *Bmp10*, *Gdf1* and *Gdf3*, and *Lefty* gene families. Anthozoan *Activin* genes cluster only with low support with bilaterian *Activin* genes. Similarly,

*Nodal*-related sequences from *A. millepora*, *A. digitifera*, *Branchiostoma floridae*, *C. pacificum*, *H. magnipapillata* and *Mnemiopsis leidyi* cluster with the bilaterian *Nodal* genes. For these pre-bilaterian *Nodal*-related genes, two scenarios are possible. One scenario is that these *Ndr* genes are pre-bilaterian Tgf-β-based inventions that acquired the *Nodal* function (as has been shown here for *H. magnipapillata*). The second scenario is that these *Ndr* sequences are orthologous to the bilaterian *Nodal* genes; however, as the branches in the backbone of the tree are deep, it is not possible to verify this possibility with sufficient support, owing to saturation and noise.

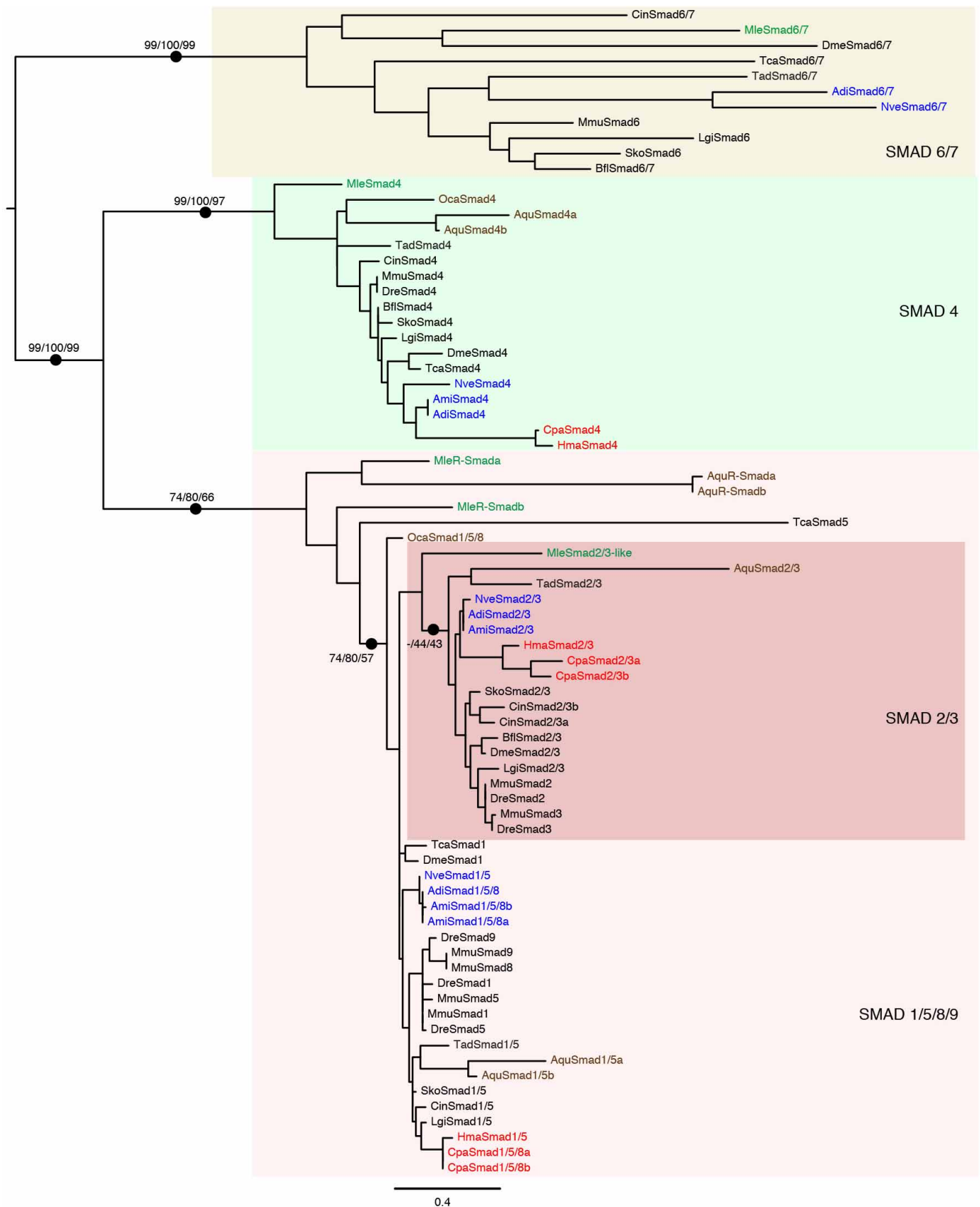




**Extended Data Figure 4 | Maximum likelihood phylogeny of the metazoan *Cerberus*, *Dan* and *Gremlin*, and *Lhx* and *Lim*, transcription factor gene families. **a**, *Cerberus*, *Dan* and *Gremlin* gene families. **b**, *Lhx* and *Lim* transcription factor gene families. *H. magnipapillata* (*Hma*) and *C. pacificum***

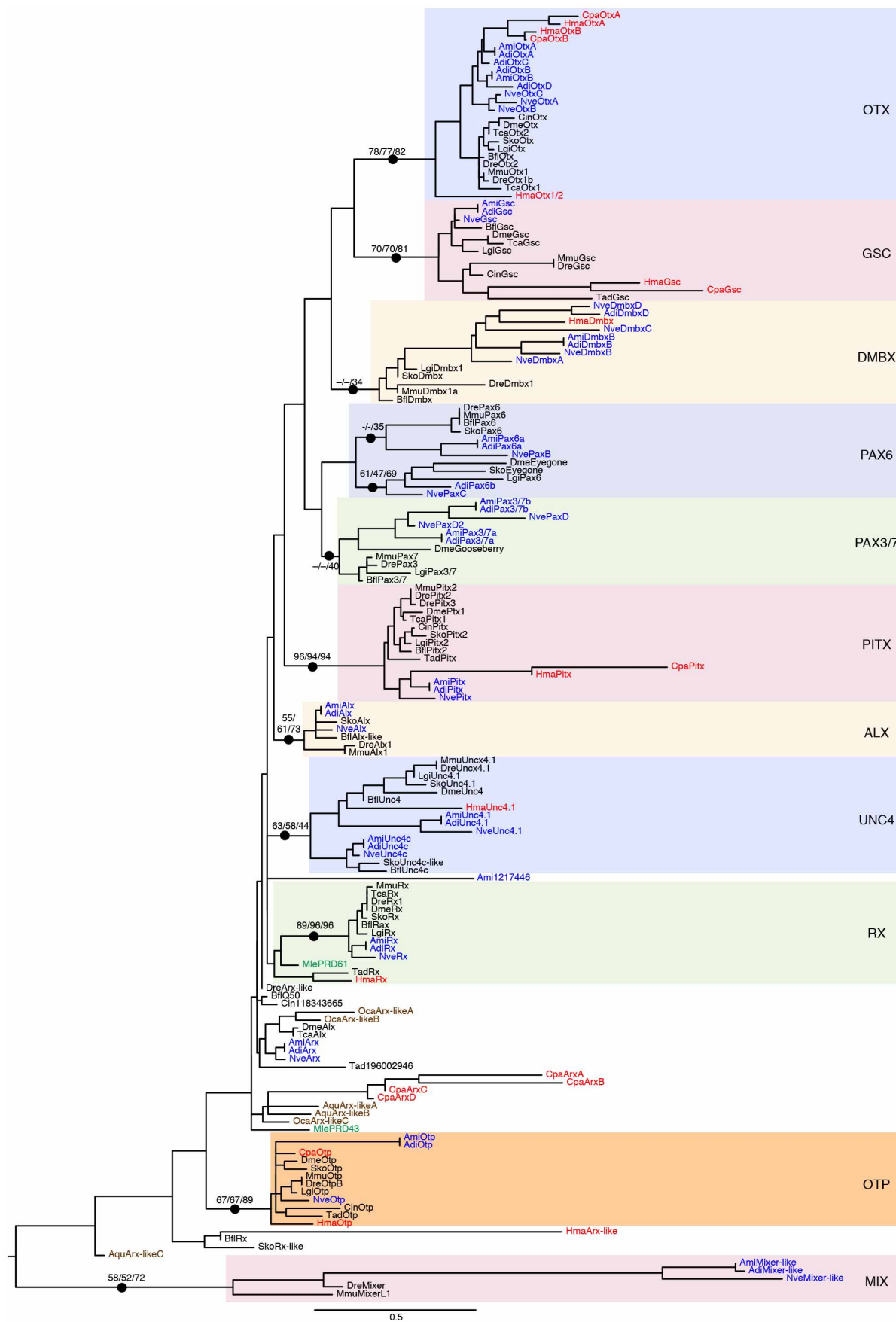
(*Cpa*) genes are shown in red, and *N. vectensis* and *Acropora* sp. genes are shown in blue. The numbers at the branches indicate the bootstrap supports from 100 BioNJ, 100 IQPNNI and 100 PhyML bootstrap trees, respectively.





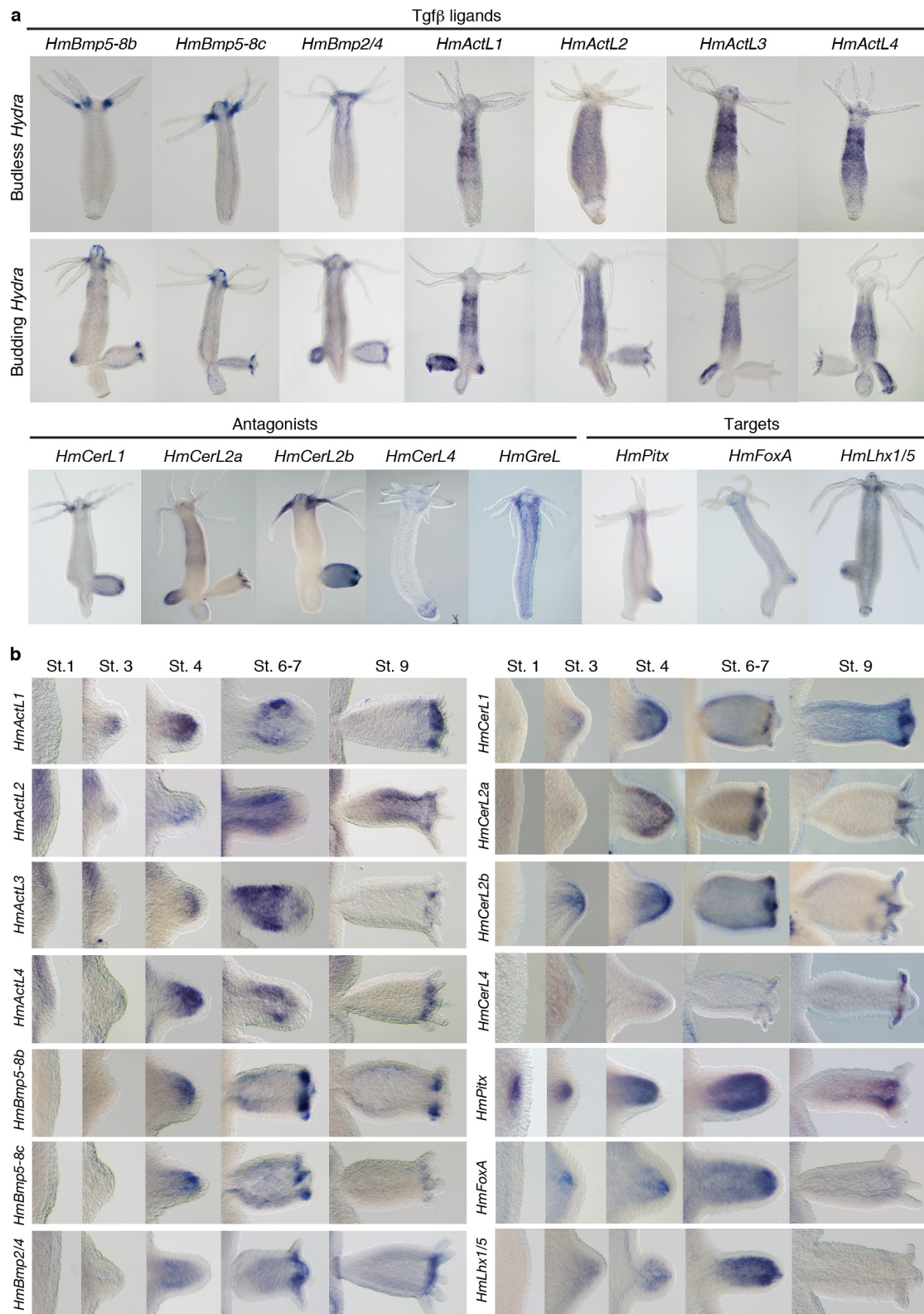
**Extended Data Figure 5 | Maximum likelihood phylogeny of the metazoan *Smad* gene family.** *H. magnipapillata* (Hma) and *C. pacificum* (Cpa) genes are shown in red, and *N. vectensis* and *Acropora* sp. genes are shown in blue.

The numbers at the branches indicate the bootstrap supports from 100 BioNJ, 100 IQPNNI and 500 PhyML bootstrap trees, respectively.



**Extended Data Figure 6 | Maximum likelihood phylogeny of the metazoan *Prd* gene superfamily of transcription factors.** *H. magnipapillata* (*Hma*) and *C. pacificum* (*Cpa*) genes are shown in red, and *N. vectensis* and *Acropora*

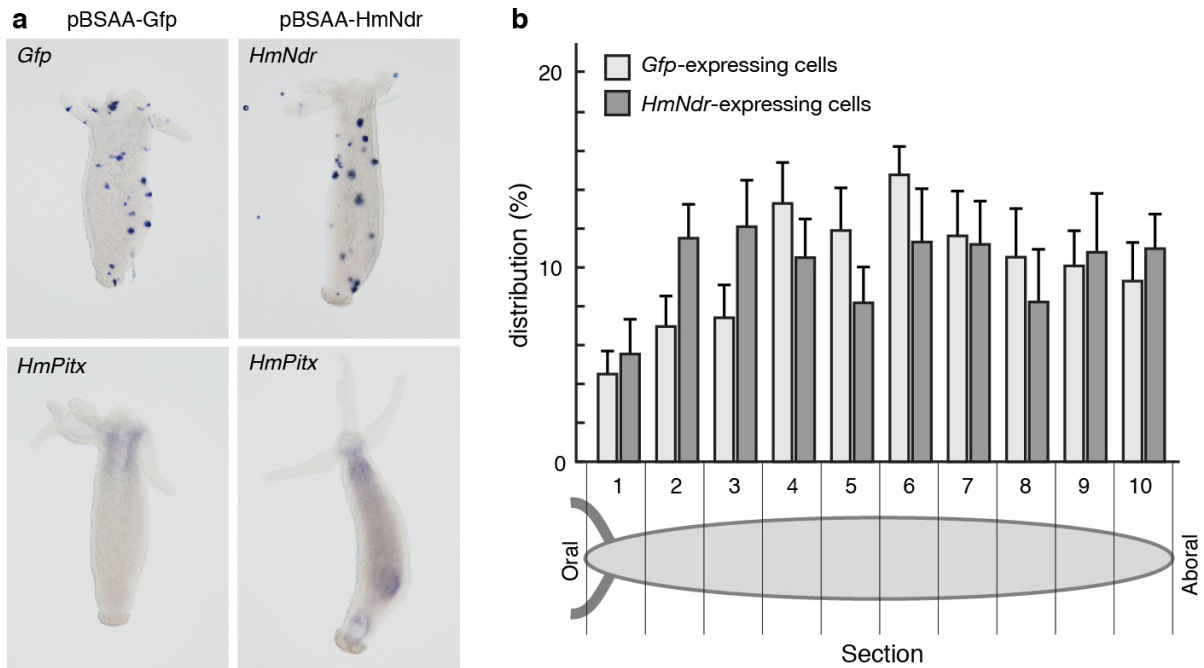
sp. genes are shown in blue. The numbers at the branches indicate the bootstrap supports from 100 BioNJ, 100 IQPNNI and 100 PhyML bootstrap trees, respectively.



**Extended Data Figure 7 | Expression patterns of Nodal-signalling-related genes in *H. magnipapillata*.** **a**, *H. magnipapillata* Bmp genes exhibit a similar pattern of expression in endodermal cells at the oral side. All Activin genes show expression in the endodermal layer in the upper half of the body column, except for the head region. *ActL2* shows a broader expression area than the other Activin genes do along the body length towards the foot region. *Cerberus-like1* (*CerL1*), *CerL2a* and *CerL2b* are expressed at a high level in the endodermal layer at the base of the tentacles in mature Hydra. *CerL4* was expressed in specific endodermal cells at the tip of the tentacles. *Gremlin-like*

(*GreL*) showed a ubiquitous expression pattern. The Nodal target genes *Pitx* and *FoxA* were expressed in the endoderm. Prominent expression was observed at the evaginating bud tip and head region (*Pitx*) or the base of the tentacles (*FoxA*). *Lhx1/5* showed clear expression at the mature and developing oral tip and peduncle. **b**, At the presumptive bud region (stage 1), *Pitx* expression was clearly demonstrated, but the other genes were under the detection limit by WISH. The first detectable expression of the *ActL1*, *CerL1*, *CerL2b* and *FoxA* genes appeared at budding stage 3.

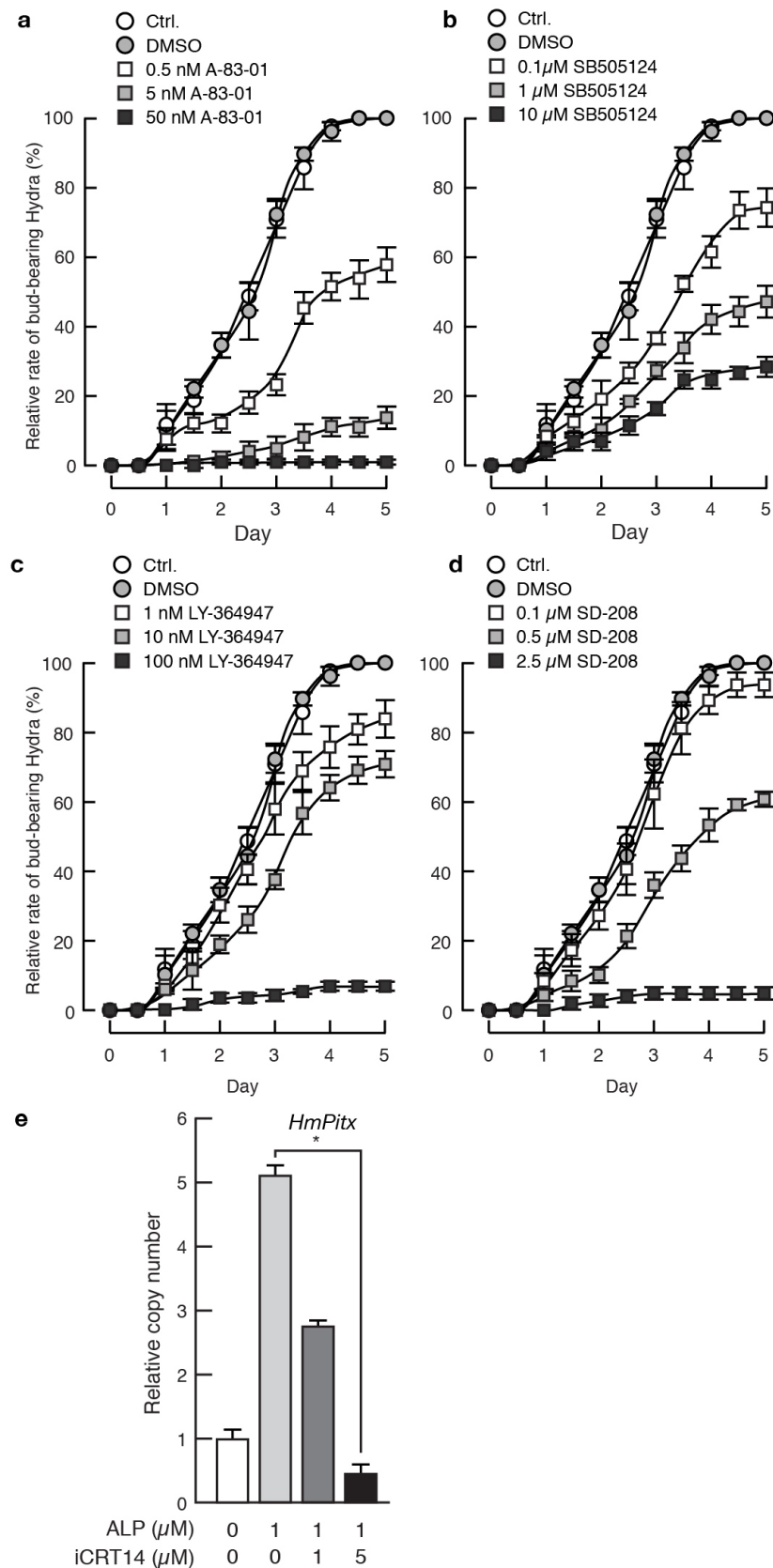




**Extended Data Figure 8 | Ectopic overexpression of *Ndr* did not induce ectopic bud formation.** **a**, Budless animals were electroporated with pBSAA-GFP (control) or pBSAA-Ndr vectors. After a 2-day incubation in hydra medium, cells expressing *GFP* or *H. magnipapillata Ndr* or *Pitx* genes were visualized by WISH. Ectopic overexpression of *Ndr* resulted in a slight increase in *Pitx* expression along the main body axis, but a new head organizer with

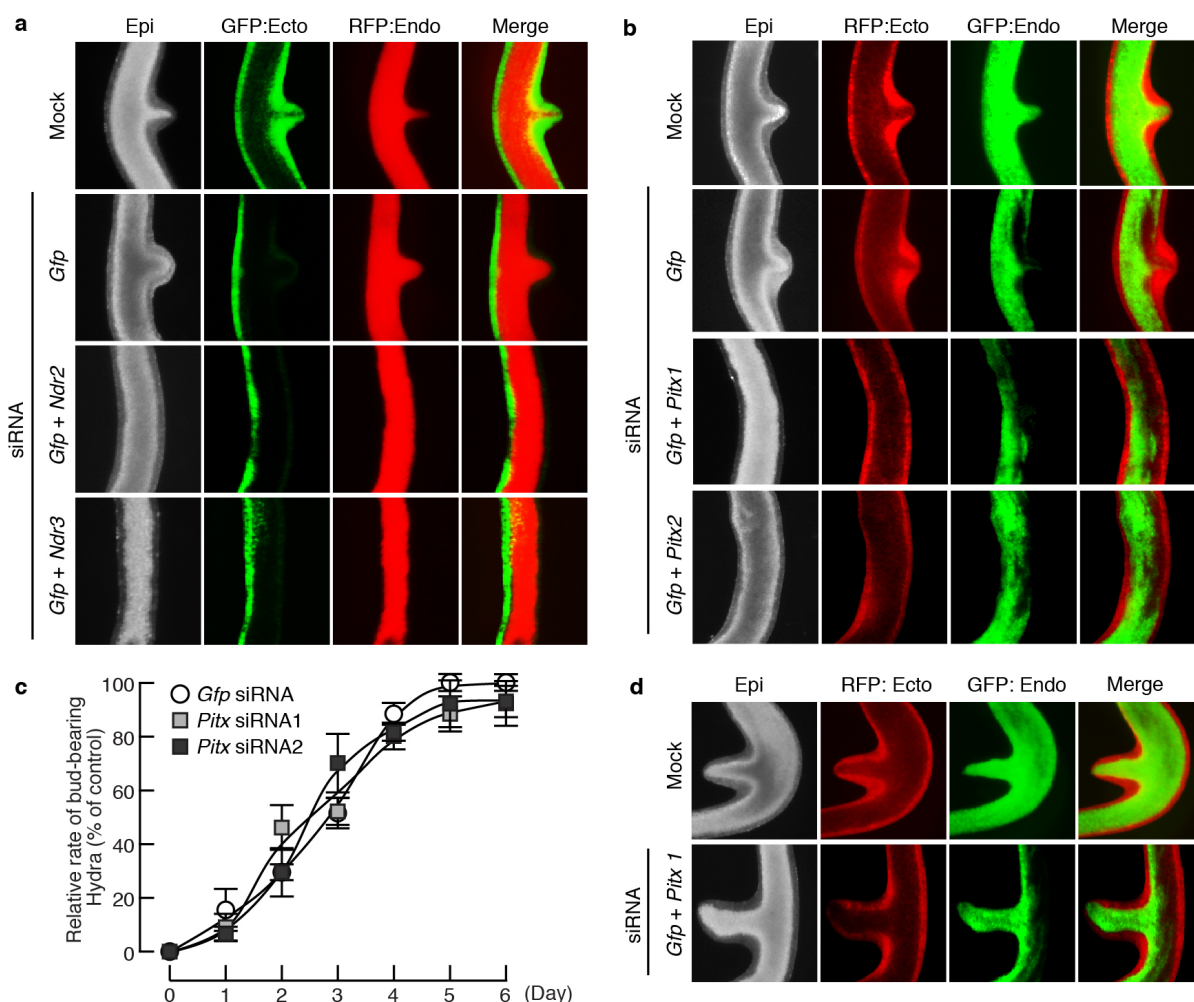
strong *Pitx* expression was established only in the budding zone.

**b**, Quantification of transfected cells demonstrated an almost equal distribution of GFP- or Ndr-expressing cells along the oral-aboral axis. This finding indicates that the whole body column region has a similar capacity to be transfected with an *Ndr*-expressing plasmid and to express an exogenous *Ndr* gene.



**Extended Data Figure 9 | Dose-dependent effect of inhibitors of Alk4/5/7 and of  $\beta$ -Catenin.** **a–d**, Budless animals were incubated in the absence or presence of specific inhibitors of Alk4/5/7: A-83-01, SB-505124, LY-364947 and SD-208. The data are presented as the mean  $\pm$  s.e.m. of four independent experiments. **e**, Hydra polyps treated with 1  $\mu$ M ALP (GSK3 inhibitor) for 16 h with or without 1 or 5  $\mu$ M iCRT14 ( $\beta$ -Catenin inhibitor). The relative

expression level of *Pitx* was examined by qPCR. The data, which were normalized to the internal standard *EF1 $\alpha$* , are presented as the ratio of the fold increase compared with the non-treated control. The data are presented as the mean  $\pm$  s.e.m. (triplicate determinations), with similar results obtained in three independent experiments. \*,  $P < 0.05$ ; one-sided *t*-test.



**Extended Data Figure 10 | Functions of *H. magnipapillata* *Ndr* and *Pitx* in bud initiation.** Gene-specific siRNAs were co-transfected with a *GFP*-specific siRNA, and bud formation from the side of transfection (*GFP*-depleted side) was monitored. In the mock-transfected hydra, *GFP* fluorescence remained intact after electroporation. As *Ndr* and *Pitx* are expressed in different germ layers (that is, in the ectoderm and endoderm, respectively), we used transgenic *Hydra* strains in which either the ectoderm or endoderm is labelled with *GFP* or red fluorescent protein (RFP), respectively<sup>35</sup>. Bud formation from the *GFP*-depleted side was significantly suppressed by co-transfection with an *Ndr*-specific siRNA (**a**) or a *Pitx*-specific siRNA (**b**). Note that the RFP fluorescence

was not affected, indicating that the decrease in the *GFP* fluorescence was not due to tissue damage caused by electroporation. The siRNAs were electroporated more effectively into the ectoderm than the endoderm, because the *GFP*-positive region of the endoderm remained more expanded at the non-transfected side (**a**, **b**). Bud formation from the side of the *GFP*-positive endoderm was not affected in *Pitx*-specific siRNA transfectants (**c**, **d**), indicating that localized expression and function of *Pitx* at the budding region is required. The data are presented as the mean  $\pm$  s.e.m. of three independent experiments.



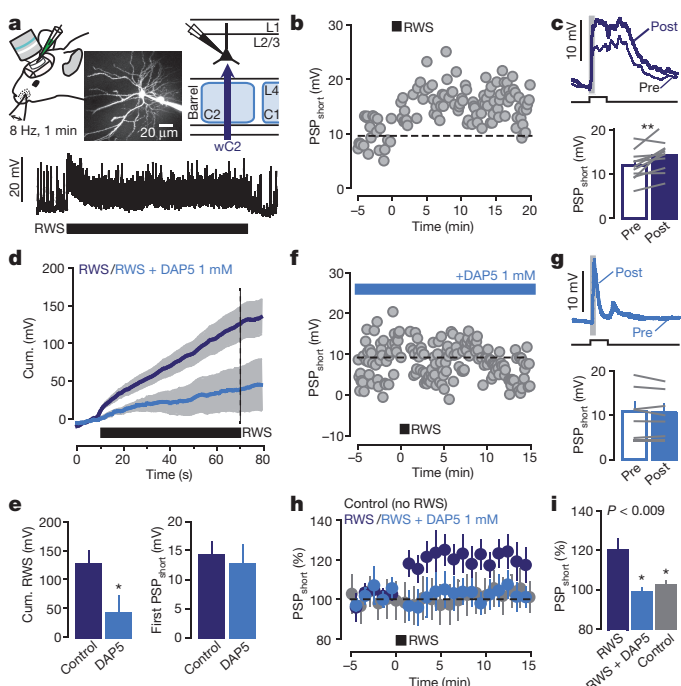
# Sensory-evoked LTP driven by dendritic plateau potentials *in vivo*

Frédéric Gambino<sup>1\*†</sup>, Stéphane Pagès<sup>1\*</sup>, Vassilis Kehayas<sup>1,2</sup>, Daniela Baptista<sup>1</sup>, Roberta Tatti<sup>1,2</sup>, Alan Carleton<sup>1</sup> & Anthony Holtmaat<sup>1</sup>

Long-term synaptic potentiation (LTP) is thought to be a key process in cortical synaptic network plasticity and memory formation<sup>1</sup>. Hebbian forms of LTP depend on strong postsynaptic depolarization, which in many models is generated by action potentials that propagate back from the soma into dendrites<sup>2,3</sup>. However, local dendritic depolarization has been shown to mediate these forms of LTP as well<sup>4,5</sup>. As pyramidal cells in supragranular layers of the somatosensory cortex spike infrequently<sup>6–8</sup>, it is unclear which of the two mechanisms prevails for those cells *in vivo*. Using whole-cell recordings in the mouse somatosensory cortex *in vivo*, we demonstrate that rhythmic sensory whisker stimulation efficiently induces synaptic LTP in layer 2/3 (L2/3) pyramidal cells in the absence of somatic spikes. The induction of LTP depended on the occurrence of NMDAR (*N*-methyl-D-aspartate receptor)-mediated long-lasting depolarizations, which bear similarities to dendritic plateau potentials<sup>9–13</sup>. In addition, we show that whisker stimuli recruit synaptic networks that originate from the posteromedial complex of the thalamus (POM). Photostimulation of channelrhodopsin-2 expressing POM neurons generated NMDAR-mediated plateau potentials, whereas the inhibition of POM activity during rhythmic whisker stimulation suppressed the generation of those potentials and prevented whisker-evoked LTP. Taken together, our data provide evidence for sensory-driven synaptic LTP *in vivo*, in the absence of somatic spiking. Instead, LTP is mediated by plateau potentials that are generated through the cooperative activity of lemniscal and paralemniscal synaptic circuitry<sup>14–16</sup>.

In most cortical synaptic LTP studies *in vivo*, strong postsynaptic depolarization was provided by action-potential-triggering somatic current injections<sup>3</sup>. To examine whether sensory stimuli can elicit LTP of synaptic inputs on L2/3 pyramidal cells in the mouse somatosensory cortex without the help of artificially triggered backpropagating action potentials, we recorded sensory-evoked postsynaptic potentials (PSPs) in the barrel cortex *in vivo*, and applied a rhythmic whisker stimulation (RWS) protocol that has been shown to enhance whisker-evoked local field potentials<sup>17</sup>. Whole-cell patch recordings were targeted to cells above the C2 barrel of urethane-anaesthetized mice<sup>6,18</sup> (Fig. 1a). Using a piezoelectric actuator, the principal whisker was deflected back and forth (100-ms deflections) for 1 min at a frequency of 8 Hz, which is within the range of frequencies at which mice sample objects<sup>15</sup>. In all cells, RWS evoked a sustained subthreshold depolarization (Fig. 1a), which was reminiscent of an evoked, NMDAR-dependent, cortical upstate<sup>19,20</sup>. None of the recorded cells displayed somatic action potentials during RWS. On average, RWS elicited a significant potentiation of subsequent single (0.1 Hz) whisker-deflection-evoked short-latency PSP amplitudes (PSP<sub>short</sub>; Fig. 1b, c; values for Figs 1–4 are provided in Supplementary Information). This LTP lasted for as long as the cells could be recorded from (at least 15 min after RWS), and correlated with the decay time of the sustained depolarization (Extended Data Fig. 1a, b). Pharmacological or hyperpolarization-mediated suppression of NMDAR conductance specifically attenuated the RWS-evoked sustained depolarization,

and prevented LTP (Fig. 1d–i, Extended Data Fig. 1 and Supplementary Note 1). RWS-induced LTP remained specific to the rhythmically stimulated synaptic pathway (Extended Data Fig. 2 and Supplementary Note 2), increased gradually over a timescale of minutes, and was occluded upon a second RWS 10 min after the first stimulation (Extended Data Fig. 3a, b). RWS and hyperpolarization did not significantly alter intrinsic cell membrane properties, and small changes thereof did not correlate with the

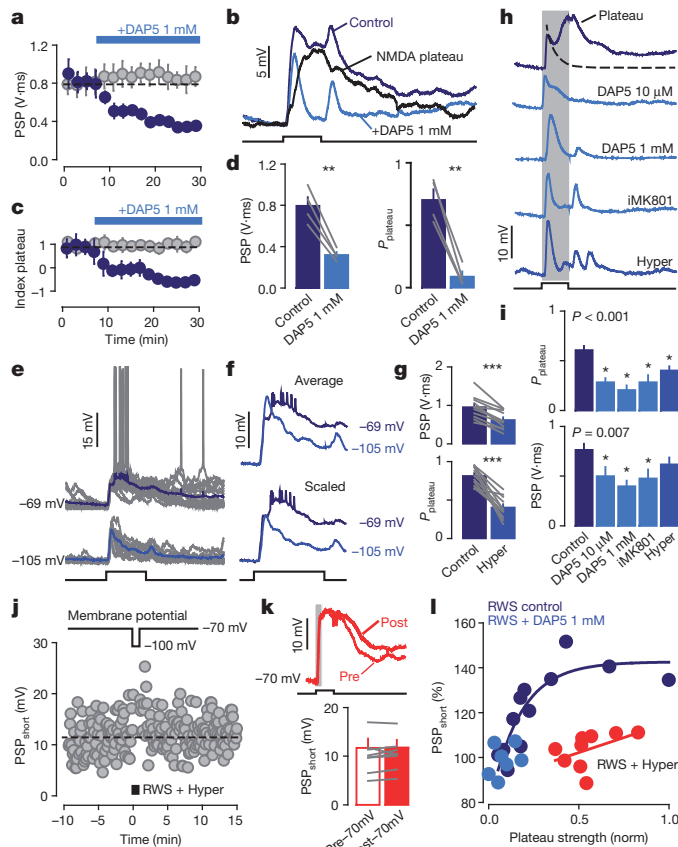


**Figure 1 | Rhythmic whisker stimulation induces LTP in L2/3 neurons.**

**a**, Top, schematic of recordings in L2/3 cells *in vivo*. PSPs and RWS are evoked by the principal whisker (wC2). Bottom, example trace of sustained depolarization induced by RWS (8 Hz for 1 min; black bar). **b**, Single-cell whisker-evoked short latency PSP amplitudes (PSP<sub>short</sub>; see **c**) before and after RWS. **c**, Top, single-cell averaged traces, pre and post RWS. Grey box, PSP<sub>short</sub> window. Bottom, mean amplitudes, pre and post RWS ( $n = 11$ ,  $**P = 0.008$ ). **d**, Cumulative RWS-induced depolarization. **e**, Mean cumulative depolarization at the end of RWS and amplitudes of the first PSP<sub>short</sub> upon RWS (control,  $n = 11$ ; +DAP5,  $n = 7$ ;  $P = 0.028$  (left) and  $P = 0.703$  (right)). **f**, Single-cell whisker-evoked PSP<sub>short</sub> amplitudes upon epidural DAP5 (1 mM). **g**, Top, single-cell averaged traces, pre and post RWS under DAP5. Bottom, mean amplitudes, pre and post RWS under DAP5 ( $n = 7$ ,  $P = 0.344$ ). **h**, Mean PSP<sub>short</sub> amplitudes without RWS, and pre and post RWS in controls and under DAP5. **i**, Mean amplitudes normalized to baseline (RWS,  $n = 11$ ; RWS + DAP5,  $n = 7$ ; Control,  $n = 7$ ;  $P < 0.009$ , one-way ANOVA;  $*P < 0.05$ , post-hoc comparisons versus RWS control). Error bars, s.e.m.; square pulse lines, whisker deflections (100 ms); grey lines between bars, pairs.

<sup>1</sup>Department of Basic Neurosciences and the Center for Neuroscience, CMU, University of Geneva, 1 rue Michel Servet, 1211 Geneva, Switzerland. <sup>2</sup>Lemanic Neuroscience Doctoral School, 1 rue Michel Servet, 1211 Geneva, Switzerland. <sup>†</sup>Present address: Institute for Interdisciplinary Neuroscience (IINS), UMR 5297 CNRS and University of Bordeaux, 146 rue Léo-Saignat, 33077 Bordeaux, France.

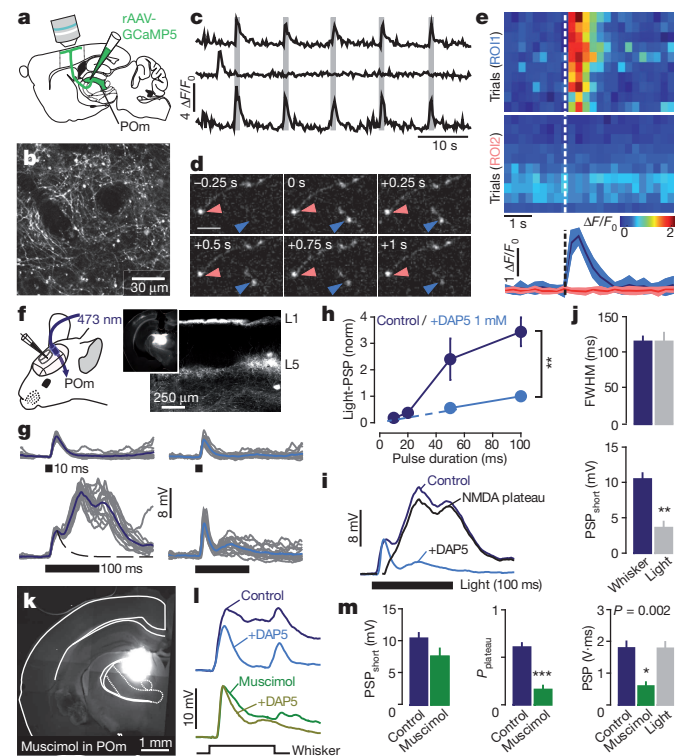
\*These authors contributed equally to this work.



**Figure 2 | Whisker deflections evoke NMDAR-mediated plateau potentials.** **a**, Mean whisker-evoked 100-ms PSP integrals in controls (grey,  $n = 7$ ), and upon epidural DAP5 (blue,  $n = 5$ ). **b**, Single-cell averaged traces ( $n = 40$  trials). NMDAR-mediated plateau: DAP5-trace subtracted from control trace. **c**, Mean Index<sub>plateau</sub> (see Extended Data Fig. 4) of same cells as in **a**. **d**, Mean PSP integrals and plateau probabilities ( $n = 5$ ,  $^{**}P = 0.004$  (left) and  $^{**}P = 0.002$  (right)). **e**, Single-cell examples of responses at two different holding potentials (grey, single trials; blue, averaged traces). **f**, Top, averaged traces from **e**. Bottom, traces normalized to PSP<sub>short</sub> in order to compensate for increased driving force, showing that hyperpolarization blocks the plateau. **g**, Mean PSP integrals and plateau probabilities ( $n = 11$ ,  $^{***}P < 0.001$ ). **h**, Single-cell averaged traces under control (dark blue) and various NMDAR conductance-reducing conditions. **i**, Comparison of mean plateau potential probabilities and integrals under various conditions (control,  $n = 44$ ; 10  $\mu$ M DAP5,  $n = 9$ ; 1 mM DAP5,  $n = 12$ ; iMK801,  $n = 10$ ; Hyper,  $n = 11$ ;  $P < 0.001$  (top) and  $P = 0.007$  (bottom), one-way ANOVA;  $^{*}P < 0.05$ , post-hoc comparisons versus control). **j**, Single-cell PSP<sub>short</sub> amplitudes recorded at resting membrane potential ( $-70$  mV) before and after RWS. The cell was hyperpolarized ( $-100$  mV) during RWS. **k**, Top, single-cell averaged traces, pre and post RWS upon hyperpolarization. Grey box, PSP<sub>short</sub> window. Bottom, mean amplitude, pre and post RWS recorded at  $-70$  mV, but upon hyperpolarization during RWS ( $n = 9$ ,  $P = 0.993$ ). **l**, Normalized plateau strength predicts the level of RWS-induced LTP in controls (nonlinear regression;  $R^2 = 0.76$ ,  $P = 0.002$ ). DAP5 blocks plateaus and LTP. Neurons with high plateau strengths fail to potentiate when hyperpolarized during RWS (PSP<sub>short</sub>, mean:  $103.6 \pm 6\%$ ,  $n = 9$ ). Error bars, s.e.m.; square pulse lines, whisker deflections (100 ms); grey lines between bars, pairs.

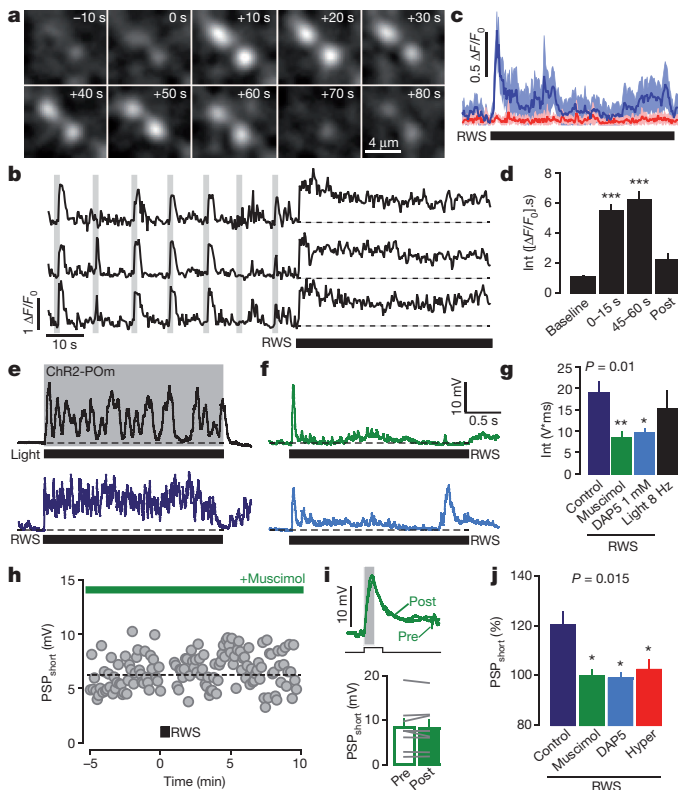
magnitude of LTP (Extended Data Fig. 3c–n). Together, the data indicate that RWS elicits bona fide synaptic LTP by evoking a cell autonomous and sustained NMDAR-dependent subthreshold depolarization.

We next investigated what may cause the sustained subthreshold depolarization upon RWS. In accordance with previous reports single whisker deflections typically evoked compound PSPs, containing short- and long-latency components, which is similar to the sparse, temporal, spiking profiles of L2/3 cells (Fig. 1c and Extended Data Fig. 4a, b)<sup>6,21–23</sup>. Short-latency PSPs were always present. Long-latency PSPs occurred with variable probabilities in different cells. They had an all-or-none, nonlinear appearance



**Figure 3 | Plateau potentials depend on POM activity.** **a**, **b**, POM neurons were transfected with adeno-associated virus (AAV)-GCaMP5G and their axons were imaged in L1 of the barrel cortex (**b**). **c**, Example traces of  $\text{Ca}^{2+}$  signals ( $\Delta F/F_0$ ) in POM axonal boutons upon whisker deflection bouts (5 at 20 Hz; grey lines). **d**, Example of a responsive (blue arrowhead) and an unresponsive (red arrowhead) axon. Scale bar, 5  $\mu$ m. **e**, Top, whisker-evoked (dashed line)  $\text{Ca}^{2+}$  signals ( $\Delta F/F_0$ ) in the boutons of **d** (rows, 10 successive trials). Bottom, mean whisker-evoked  $\text{Ca}^{2+}$  signals in the two boutons. **f**, Photostimulation of ChR2-expressing POM neurons using an optical fibre, and ChR2 expression profiles in the thalamus (inset) and cortex. **g**, Single-cell examples of photostimulus-evoked PSPs (in L2/3 cells) before (left) and after DAP5 (right). Grey, single trials; blue, averaged traces; black bar, photostimulus. **h**, Photostimulus-evoked PSP integrals as a function of stimulus duration (normalized to the maximal integral upon a 100-ms pulse + DAP5 (control,  $n = 7$ ; DAP5,  $n = 5$ ;  $^{**}P = 0.002$ )). **i**, Averaged 100-ms light-pulse-evoked PSPs from **g**. NMDA-plateau: DAP5-trace subtracted from control trace. **j**, Comparison of 100-ms light pulse-evoked and whisker-evoked PSPs (whisker,  $n = 33$ ; light,  $n = 11$ ;  $P = 0.918$  (top);  $^{**}P < 0.01$  (bottom)). **k**, Coronal brain section with fluorescent muscimol in the POM. **l**, Single-cell averaged traces of 100-ms whisker-evoked PSPs in controls and after POM inactivation, with and without DAP5. **m**, Effects of POM inactivation on PSP<sub>short</sub> (control,  $n = 33$ ; muscimol,  $n = 9$ ;  $P = 0.095$ ), plateau potential probabilities (control,  $n = 44$ ; muscimol,  $n = 9$ ;  $^{***}P < 0.001$ , Mann–Whitney U-test), and PSP integrals (over 300 ms; control,  $n = 33$ ; muscimol,  $n = 9$ ; 100 ms light,  $n = 13$ ;  $P = 0.002$ , one-way ANOVA, and  $^{*}P < 0.05$ , post-hoc comparisons versus control). Error bars represent s.e.m.

and were selectively attenuated upon an NMDAR block (Fig. 2a–d and Extended Data Fig. 4a–g). This indicates that long-latency PSPs share similarities with plateau potentials<sup>9–13</sup>. We termed them accordingly. Importantly, plateau potentials, but not short-latency PSPs, were also significantly attenuated upon a cell-autonomous suppression of NMDAR conductance using artificial hyperpolarization ( $-100$  mV), or intracellular MK801 (1 mM) (Fig. 2e–i and Extended Data Fig. 4g). Conversely, when holding the neurons at a slightly depolarized state, or in some occasions at resting states, these potentials evoked spikes (Fig. 2e). This suggests that whisker-evoked long-latency spiking under normal conditions is associated with the occurrence of NMDAR-dependent plateau potentials (Extended Data Fig. 4b)<sup>10,13,21,24</sup>. The occurrence of plateau potentials was not critically dependent on the 100-ms duration of the whisker deflections (Extended Data Fig. 4h–j).



**Figure 4 | RWS-evoked LTP depends on POM-mediated plateau potentials.** **a**, Example of POM axonal bouton  $\text{Ca}^{2+}$  dynamics ( $\Delta F/F_0$ ) upon RWS. **b**,  $\text{Ca}^{2+}$  signals in POM boutons upon whisker-deflection bouts (grey lines) and RWS. **c**, Mean RWS-induced  $\text{Ca}^{2+}$  signals in the 25 most active (blue) and least active (red) boutons (see Extended Data Fig. 6). Shaded area, s.d. **d**, Integrated responses before, during and after RWS ( $n = 765$ ; \*\*\* $P < 0.001$ ). **e**, Example of the membrane depolarization (L2/3 cell) upon rhythmic photostimulation of POM (20 flashes at 8 Hz) or RWS (20 stimuli at 8 Hz). Black bar, stimulus period. **f**, RWS-induced membrane depolarization upon muscimol-POM inactivation or DAP5. **g**, Mean integrated membrane potentials (over 2.5 s) (control,  $n = 14$ ; muscimol,  $n = 4$ ; DAP5,  $n = 15$ ; light 8 Hz,  $n = 5$ ;  $P = 0.01$ , one-way ANOVA; \* $P < 0.05$  and \*\* $P < 0.01$ , post-hoc comparisons versus control). **h**, Single-cell whisker-evoked  $\text{PSP}_{\text{short}}$  amplitudes before and after RWS upon muscimol-mediated POM inactivation. **i**, Top, single-cell averaged traces, pre and post RWS. Grey line,  $\text{PSP}_{\text{short}}$  window. Square pulse line, whisker deflection (100 ms). Bottom, mean amplitudes pre and post RWS upon POM inactivation ( $n = 9$ ,  $P = 0.972$ ). Grey lines between bars, pairs. **j**, Comparison of normalized  $\text{PSP}_{\text{short}}$  amplitudes upon RWS in controls, and after a POM or NMDAR block (control,  $n = 11$ ; muscimol,  $n = 9$ ; DAP5,  $n = 7$ ; Hyper,  $n = 9$ ;  $P = 0.015$ , one-way ANOVA, and \* $P < 0.05$ , post-hoc comparisons versus control). Error bars, s.e.m.

NMDAR-mediated plateau potentials or summing NMDA spikes have been observed in various cell types in the somatosensory cortex *in vivo*<sup>10,11,13</sup>, where they are characterized by local and spreading dendritic  $\text{Ca}^{2+}$  transients. We confirmed that under our conditions too, single whisker deflections evoke  $\text{Ca}^{2+}$  transients in spines and dendritic shafts, both under wakefulness (data not shown) and anaesthesia. Local and large-scale events occurred at various positions in the dendritic tree (Extended Data Fig. 5a–e and Supplementary Note 3). These responses were diminished upon an NMDAR block, and increased upon RWS (Extended Data Fig. 5f–j), similar to the plateau potentials and sustained depolarization in our whole-cell recordings. Together, this indicates that single-whisker-deflection-mediated plateaus and the RWS-mediated sustained depolarization in our recordings are likely to be supported by NMDAR-mediated dendritic  $\text{Ca}^{2+}$  events<sup>10,11,13,25</sup>.

The plateau strength of neurons (that is, the product of the probability for a whisker deflection to elicit a plateau potential and the average integrated depolarization) correlated with the magnitude of RWS-induced

LTP (Fig. 2l). Neurons bearing high plateau strengths could not be potentiated when they were hyperpolarized during RWS (Fig. 2j–l). This strongly suggests that the plateau potentials are essential for RWS-induced LTP.

Next, we questioned what could be the synaptic source of NMDAR-mediated plateau potentials. Previous studies indicate that in the somatosensory cortex they are generated by coincident activity of segregated excitatory synaptic pathways<sup>10,11,13</sup>. We reasoned that in anaesthetized mice plateau potentials may depend on co-activation of intracolumnar lemniscal and thalamocortical paralemniscal pathways<sup>10,11</sup>. Paralemniscal input to L2/3 cells may be directly or indirectly provided by thalamic POM efferents that target pyramidal cell dendrites in L5A and L1 (refs 14–16, 26–29). This system is involved in the kinematics of whisking<sup>15</sup> and may provide feedback information to pyramidal cells<sup>16</sup>.

To test whether POM efferents are activated by passive whisker stimuli<sup>15,28,29</sup>, we expressed the genetically encoded  $\text{Ca}^{2+}$  indicator GCaMP5G in POM neurons (Fig. 3a, b), and imaged responses in their projections to somatosensory cortex L1 through a cranial window. Whisker deflections evoked  $\text{Ca}^{2+}$  transients with various latencies in a substantial portion of the terminals in awake and anaesthetized mice (Fig. 3c–e and Extended Data Fig. 6). The shortest response times matched the latencies of the plateau potentials. Furthermore, the pattern of activation was widespread and did not remain limited to the whisker's home barrel column (Extended Data Fig. 7a–c and Supplementary Note 4). Interestingly, and in contrast to short-latency PSPs, the plateau potentials were not selective for the principal whisker, which supports the possibility that the POM thalamocortical circuitry is indeed involved in generating them (Extended Data Fig. 7d, e).

To test this directly we expressed the recombinant light-gated ion channel channelrhodopsin-2-Venus (ChR2-Venus) in the POM and recorded photostimulus-evoked PSPs in L2/3 cells (Fig. 3f, g and Extended Data Fig. 8a–c and Supplementary Note 5). POM neurons were stimulated using an optical fibre that was guided by a stereotactically implanted cannula (Fig. 3f). Short square light pulses (10–20 ms) produced small-amplitude PSPs, longer pulses (50–100 ms) produced plateaus (Fig. 3g–h). This indicates that a strong activation of POM thalamic nuclei may generate plateau potentials by itself, through monosynaptic inputs (Extended Data Fig. 8d, e and Supplementary note 5) or, more probably, through the large-scale recruitment of cortical paralemniscal synaptic networks that project to L2/3 cells<sup>14,27,28</sup>. This suggests that the POM-associated synaptic circuitry mediates whisker-evoked plateau potentials. Indeed, analogous to whisker-evoked plateaus, photostimulus-evoked plateaus were eliminated upon an NMDAR block (Fig. 3g, h). The full width at half maximum of the NMDAR-mediated plateaus was equal between whisker and 100-ms light stimuli (Fig. 3i, j; for example, compare with Fig. 2b). Importantly, for these photostimuli the peak amplitude of the short-latency PSPs was significantly lower than that of whisker-evoked PSPs (Fig. 3j), confirming that photostimuli did not generate plateau potentials through a large-scale recruitment of lemniscal synaptic pathways.

To test further the role of this paralemniscal circuitry, we specifically suppressed POM activity using the GABA-A-R selective agonist muscimol (Fig. 3k, Extended Data Fig. 9a and Supplementary Note 6). This did not affect whisker-evoked short-latency PSPs, but greatly reduced whisker-evoked plateau potential probabilities (Fig. 3l, m). The inhibition of POM activity was sufficient to eliminate most of the evoked NMDAR-mediated plateaus, since an additional NMDAR block did not further reduce them (Fig. 3l).

The role of the POM in the generation of plateau potentials suggests that it is also involved in the production of the RWS-evoked sustained depolarization. Indeed, RWS evoked sustained  $\text{Ca}^{2+}$  transients in a portion of POM-derived axonal boutons (Fig. 4a–d and Extended Data Fig. 6e, f). Rhythmic photostimulation of ChR2-expressing POM neurons evoked a sustained depolarization similar to RWS (Fig. 4e, g). Conversely, RWS failed to evoke a sustained depolarization upon muscimol-mediated suppression of POM activity, similar to the effect of DAP5 (Fig. 4f, g). This suggests that POM activity evokes multiple plateaus during RWS. To test the causal relationship between plateau potentials, RWS-evoked sustained



depolarization, and RWS-induced LTP without blocking NMDARs, we applied RWS to muscimol-injected mice. In these mice RWS failed to induce LTP (Fig. 4h, i), similar to the effect of suppressing NMDAR conductance (Fig. 4j). This strongly suggests that plateau potentials, driven by the co-activation of POM-associated synaptic circuitry facilitate whisker-evoked LTP.

In summary, our data provide evidence for sensory-evoked LTP that is independent of somatic spikes, displaying similarities to *in vitro* experiments in hippocampus<sup>30</sup>. *In vivo*, the occurrence of LTP in the absence of somatic spikes may be an important mechanism to strengthen synapses between weakly connected neurons, without the necessity for sensory inputs to first elicit a sufficient number of action potentials. Since L2/3 neurons in the somatosensory cortex normally spike sparsely or infrequently<sup>8</sup>, this mechanism may also prevent neurons from losing synaptic input due to spike-timing-dependent long-term depression (LTD)-like processes<sup>3</sup>.

We also found that sensory-evoked LTP may depend on paralemniscal synaptic inputs that originate from the POM of the thalamus. This circuitry provides contextual or predictive information for external sensory stimuli that feed forward through lemniscal pathways<sup>16</sup>. Our data suggest that the repeated coincident activity of this feedback circuitry may increase L2/3 neurons' sensitivity to future sensory stimuli. It is likely that during wakefulness dendritic plateau potentials that are mediated by inputs from motor cortex<sup>11</sup> also function to facilitate this—perhaps Hebbian—form of LTP (Supplementary Note 7).

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 3 February; accepted 7 July 2014.**

**Published online 31 August 2014.**

- Bliss, T. V. P. & Collingridge, G. L. A synaptic model of memory: long-term potentiation in the hippocampus. *Nature* **361**, 31–39 (1993).
- Markram, H., Luebke, J., Frotscher, M. & Sakmann, B. Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. *Science* **275**, 213–215 (1997).
- Feldman, D. E. The spike-timing dependence of plasticity. *Neuron* **75**, 556–571 (2012).
- Golding, N. L., Staff, N. P. & Spruston, N. Dendritic spikes as a mechanism for cooperative long-term potentiation. *Nature* **418**, 326–331 (2002).
- Lisman, J. & Spruston, N. Postsynaptic depolarization requirements for LTP and LTD: a critique of spike timing-dependent plasticity. *Nature Neurosci.* **8**, 839–841 (2005).
- Brecht, M., Roth, A. & Sakmann, B. Dynamic receptive fields of reconstructed pyramidal cells in layers 3 and 2 of rat somatosensory barrel cortex. *J. Physiol. (Lond.)* **553**, 243–265 (2003).
- Poulet, J. F. & Petersen, C. C. Internal brain state regulates membrane potential synchrony in barrel cortex of behaving mice. *Nature* **454**, 881–885 (2008).
- O'Connor, D. H., Peron, S. P., Huber, D. & Svoboda, K. Neural activity in barrel cortex underlying vibrissa-based object localization in mice. *Neuron* **67**, 1048–1061 (2010).
- Antic, S. D., Zhou, W. L., Moore, A. R., Short, S. M. & Ikonou, K. D. The decade of the dendritic NMDA spike. *J. Neurosci. Res.* **88**, 2991–3001 (2010).
- Lavzin, M., Rapoport, S., Polsky, A., Garion, L. & Schiller, J. Nonlinear dendritic processing determines angular tuning of barrel cortex neurons *in vivo*. *Nature* **490**, 397–401 (2012).
- Xu, N. L. *et al.* Nonlinear dendritic integration of sensory and motor input during an active sensing task. *Nature* **492**, 247–251 (2012).
- Major, G., Larkum, M. E. & Schiller, J. Active properties of neocortical pyramidal neuron dendrites. *Annu. Rev. Neurosci.* **36**, 1–24 (2013).
- Palmer, L. M. *et al.* NMDA spikes enhance action potential generation during sensory input. *Nature Neurosci.* **17**, 383–390 (2014).
- Bureau, I., von Saint Paul, F. & Svoboda, K. Interdigitated paralemniscal and lemniscal pathways in the mouse barrel cortex. *PLoS Biol.* **4**, e382 (2006).
- Diamond, M. E., von Heimendahl, M., Knutsen, P. M., Kleinfeld, D. & Ahissar, E. 'Where' and 'what' in the whisker sensorimotor system. *Nature Rev. Neurosci.* **9**, 601–612 (2008).
- Larkum, M. A cellular mechanism for cortical associations: an organizing principle for the cerebral cortex. *Trends Neurosci.* **36**, 141–151 (2013).
- Megevang, P. *et al.* Long-term plasticity in mouse sensorimotor circuits after rhythmic whisker stimulation. *J. Neurosci.* **29**, 5326–5335 (2009).
- Gambino, F. & Holtmaat, A. Spike-timing-dependent potentiation of sensory surround in the somatosensory cortex is facilitated by deprivation-mediated disinhibition. *Neuron* **75**, 490–502 (2012).
- Poulet, J. F., Fernandez, L. M., Crochet, S. & Petersen, C. C. Thalamic control of cortical states. *Nature Neurosci.* **15**, 370–372 (2012).
- Favero, M. & Castro-Alamancos, M. A. Synaptic cooperativity regulates persistent network activity in neocortex. *J. Neurosci.* **33**, 3151–3163 (2013).
- Armstrong-James, M., Welker, E. & Callahan, C. A. The contribution of NMDA and non-NMDA receptors to fast and slow transmission of sensory information in the rat SI barrel cortex. *J. Neurosci.* **13**, 2149–2160 (1993).
- Petersen, C. C., Grinvald, A. & Sakmann, B. Spatiotemporal dynamics of sensory responses in layer 2/3 of rat barrel cortex measured *in vivo* by voltage-sensitive dye imaging combined with whole-cell voltage recordings and neuron reconstructions. *J. Neurosci.* **23**, 1298–1309 (2003).
- Wilent, W. B. & Contreras, D. Synaptic responses to whisker deflections in rat barrel cortex as a function of cortical layer and stimulus intensity. *J. Neurosci.* **24**, 3985–3998 (2004).
- Rema, V., Armstrong-James, M. & Ebner, F. F. Experience-dependent plasticity of adult rat S1 cortex requires local NMDA receptor activation. *J. Neurosci.* **18**, 10196–10206 (1998).
- Grienberger, C., Chen, X. & Konnerth, A. NMDA receptor-dependent multidendritic Ca<sup>2+</sup> spikes required for hippocampal burst firing *in vivo*. *Neuron* **81**, 1274–1281 (2014).
- Deschênes, M., Veinante, P. & Zhang, Z. W. The organization of corticothalamic projections: reciprocity versus parity. *Brain Res. Brain Res. Rev.* **28**, 286–308 (1998).
- Petreanu, L., Mao, T., Sternson, S. M. & Svoboda, K. The subcellular organization of neocortical excitatory connections. *Nature* **457**, 1142–1145 (2009).
- Feldmeyer, D. Excitatory neuronal connectivity in the barrel cortex. *Front. Neuroanat.* **6**, 24 (2012).
- Diamond, M. E., Armstrong-James, M., Budway, M. J. & Ebner, F. F. Somatic sensory responses in the rostral sector of the posterior group (POM) and in the ventral posterior medial nucleus (VPM) of the rat thalamus: dependence on the barrel field cortex. *J. Comp. Neurol.* **319**, 66–84 (1992).
- Hardie, J. & Spruston, N. Synaptic depolarization is more effective than back-propagating action potentials during induction of associative long-term potentiation in hippocampal pyramidal neurons. *J. Neurosci.* **29**, 3233–3241 (2009).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** We thank E. Ahissar and T. Oram for advice on the muscimol experiments. We appreciate C. Lüscher's comments on our manuscript. We thank L. L. Looger and D. Kim of the GENIE project, and K. Svoboda at the Janelia Farm Research Campus (HHMI) for distributing the GCaMP5G, GCaMP6S and ChR2 vectors, respectively. This work was supported by the Swiss National Science Foundation (grants 31003A\_120685, 31003A\_135631 and CRSI33\_127289 to A.H.; 31003A\_153410 to A.C.), the National Centre of Competence in Research (NCCR) SYNAPSY financed by the Swiss National Science Foundation (51AU40\_125759), the International Foundation for Research on Paraplegia, and the Hans Wilsdorf Foundation. V.K. was supported by SyMBaD (EU FP7-PEOPLE-ITN Marie Curie, grant 238608).

**Author Contributions** F.G. performed the *in vivo* electrophysiology experiments; S.P. performed the Ca<sup>2+</sup> imaging experiments; V.K. and F.G. performed the photostimulation experiments; D.B. and V.K. performed thalamocortical projection analyses; R.T. and F.G. performed *in vitro* electrophysiology experiments; A.C. and A.H. provided equipment and technical expertise; F.G., S.P., V.K. and A.H. conceived the studies; A.H. supervised the research; A.H., F.G. and S.P. wrote the manuscript with help from V.K.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.H. ([anthony.holtmaat@unige.ch](mailto:anthony.holtmaat@unige.ch)).



## METHODS

All procedures were carried out in accordance with protocols approved by the ethics committee of the University of Geneva and the authorities of the Canton of Geneva.

**Surgery and intrinsic optical imaging.** Three- to four-week-old male C57BL/6 mice were used. Anaesthesia was induced using isoflurane (4% with  $\sim 0.5 \text{ l min}^{-1} \text{ O}_2$ ) and then continued using an intraperitoneal (i.p.) injection of urethane ( $1.5 \text{ g kg}^{-1}$ , in lactated ringer solution containing (in mM) 102 NaCl, 28 Na-L-lactate, 4 KCl, 1.5  $\text{CaCl}_2$ ). Body temperature was maintained at  $37^\circ\text{C}$  by a feedback-controlled heating pad (FHC). Eye ointment was applied to prevent dehydration. In accordance with Swiss Federal laws, analgesia was provided by local application of lidocaine (1%) and i.p. injection of buprenorphine (Temgesic,  $0.05 \text{ mg kg}^{-1}$ ). The skin was disinfected with ethanol 70% and betadine, and a custom-made plastic chamber was attached to the skull above the barrel cortex using dental acrylic and dental cement (Jet Repair Acrylic, Lang Dental Manufacturing). The chamber was filled with sterile cortex buffer (in mM: 125 NaCl, 5 KCl, 10 glucose, 10 HEPES, 2  $\text{CaCl}_2$  and 2  $\text{MgSO}_4$  (pH 7.4)) and sealed with a glass coverslip.

Intrinsic optical signals were imaged as described previously<sup>18</sup>, through the intact skull using a light guide system with a 700-nm (bandwidth of 20 nm) interference filter and a stable 100-W halogen light source. An image of the surface vascular pattern was taken using green light (546-nm interference filter) at the end of each imaging session. Images were acquired using the Imager 3001F (Optical Imaging, Mountainside, NJ) equipped with a large spatial  $256 \times 256$  array, fast readout, and low read noise charge-coupled device (CCD) camera. The size of imaged area was adjusted by using a combination of two lenses with different focal distances (Nikon 50 mm, bottom lens, 135 mm, upper lens, f2.0; total magnification 2.7). The CCD camera was focused on a plane 300  $\mu\text{m}$  below the skull surface. Responses were visualized by dividing the stimulus signal by the baseline signal, using the built-in Imager 3001F analysis program (Optical Imaging, Mountainside, NJ). Signals were analysed further using a custom routine in Matlab as described previously<sup>18</sup>.

**In vivo whole-cell recordings.** After imaging, adequate anaesthesia was assessed (absence of toe pinch reflexes, corneal reflexes, and vibrissae movements) and prolonged by supplementary urethane ( $0.15 \text{ g kg}^{-1}$ , i.p.) if necessary. A small  $\sim 1 \times 1 \text{ mm}$  craniotomy (centred above the C2 whisker maximum intrinsic optical signal response) was made using a pneumatic dental drill. The dura was left intact. Whole-cell patch-clamp recordings of L2/3 pyramidal neurons were obtained as described previously<sup>18,31</sup>. High positive pressure (200–300 mbar) was applied to the pipette (5–8  $\text{M}\Omega$ ) to prevent tip occlusion. After passing the pia the positive pressure was immediately reduced to prevent cortical damage. The pipette was then advanced in 1- $\mu\text{m}$  steps, and pipette resistance was monitored in the conventional voltage clamp configuration. When the pipette resistance suddenly increased, positive pressure was relieved to obtain a 3–5-G $\Omega$  seal. After break-in,  $V_m$  was measured, and dialysis was allowed to occur for at least 5 min before deflecting the whisker. Data were acquired using a Multiclamp 700B Amplifier (Molecular Devices), and digitized at 10 kHz (National Instruments), using Matlab-based Ephus software (<http://research.janelia.org/labs/display/ephus>; The Janelia Farm Research Center). Offline analysis was performed using custom routines written in IGOR Pro (WaveMetrics).

Current-clamp recordings were made using a potassium-based internal solution (in mM: 135 potassium gluconate, 4 KCl, 10 HEPES, 10 Na2-phosphocreatine, 4 Mg-ATP, 0.3 Na-GTP and 25  $\mu\text{M}$  AlexaFluor 488 hydrazide (Invitrogen), pH adjusted to 7.25 with KOH, 285 mOsm). Series and input resistance were monitored with a 100-ms long-lasting hyperpolarizing square pulse 400 ms before each whisker deflection and extracted offline by using a double-exponential fit. Recordings were discarded if the change in these parameters was larger than 30%. The bridge was usually not balanced, and liquid junction potential was not corrected. Traces were analysed as described previously<sup>18</sup>. For the action-potential analysis in Extended Data Fig. 3, action potentials were time-aligned to their respective threshold. The action-potential threshold was computed as the minimal membrane potential value at the time corresponding to the peak of the third derivative of the membrane potential, similar to methods described previously<sup>32</sup>.

PSPs were evoked by back and forth deflection of the whisker (100 ms, 0.133 Hz) using a glass capillary 4 mm away from the skin attached to a piezoelectric ceramic actuators (PL-series PICMA, Physik Instrumente). The voltage applied to the actuator was set to evoke a whisker displacement of 0.6 mm with a ramp of 7–8 ms (for details see ref. 18). The C1 and C2 whiskers were independently deflected by different piezoelectric elements.

DAP5 (1 mM or 10  $\mu\text{M}$ , Tocris) was topically applied to the dura mater, either prior or during the whole-cell recordings. Epidural application of 0.1–1 mM DAP5 has been shown to leave spiking in the thalamus intact and only minimally impact cortical L4 after 6 h of superfusion; 10  $\mu\text{M}$  is entirely ineffective in suppressing spikes in L4 (ref. 24). MK-801 (1 mM, Tocris) was included in the internal solution. After break-in we waited for at least 5 min to let MK-801 diffuse into the cells. Typically, the effects of MK-801 became visible over the time course of 4–5 min, and remained

stable thereafter. This indicates that the effects were largely due to cell autonomous blockage of NMDARs<sup>10</sup>.

**Injection of fluorescent muscimol into the POM.** Mice were anaesthetized as described above. Analgesia was provided by local application of lidocaine and an i.p. injection of buprenorphine. A burr hole was made to stereotactically inject fluorescent muscimol (Bodipy-TMR-X, 500  $\mu\text{M}$  in cortex buffer with 5% DMSO, Invitrogen) in the POM. The caudal sector of the POM that mainly projects to L1 of S1 (ref. 33) was specifically targeted using the following stereotaxic coordinates: rostrocaudal (RC),  $-2.00 \text{ mm}$ ; mediolateral (ML),  $-1.20 \text{ mm}$ ; dorsoventral (DV),  $-3.00 \text{ mm}$  from the bregma (Extended Data Fig. 9). Glass pipettes (Wiretrol, Drummond) were pulled, back-filled with mineral oil, and front-loaded with the muscimol solution (100–150 nl were delivered ( $20 \text{ nl min}^{-1}$ ) using an oil hydraulic manipulator system (MMO-220A, Narishige)). For controls, the same volume of fluorescent muscimol was injected in thalamic structures that are not directly involved in somatosensory processing (Extended Data Fig. 9). The craniotomy was then covered with Kwik-Cast (WPI) and mice were prepared for intrinsic optical imaging and whole-cell recordings as described above. To achieve a maximal suppression of neuronal activity, patch-clamp recordings were performed at least one hour but no longer than 4 h after the injection. After completion of the experiment, mice were transcardially perfused with 4% paraformaldehyde in PBS (PFA), their brains extracted and post-fixed in PFA overnight. Coronal brain sections (100  $\mu\text{m}$ ) were made to confirm the site and spread of injections (Extended Data Fig. 9).

**Virus injections and cranial windows.** For virus injections we used either pups between postnatal days 12 and 15 (P12–P15) or adults ( $>4$  weeks). In adults, anaesthesia was induced using isoflurane (4% with  $\sim 0.5 \text{ l min}^{-1} \text{ O}_2$ ) and then continued using an i.p. injection of a mixture containing medetomidin (Dorbene,  $0.2 \text{ mg kg}^{-1}$ ), midazolam (Dormicum,  $5 \text{ mg kg}^{-1}$ ) and fentanyl (Duragesic,  $0.05 \text{ mg kg}^{-1}$ ) in sterile NaCl 0.9% (MMF-mix). To prevent potential inflammation, salivary excretions or bradycardia, carprofen (Rimadyl,  $5 \text{ mg kg}^{-1}$ ) and glycopyrrolate (Robinul,  $0.01 \text{ mg kg}^{-1}$ ) were injected subcutaneously (s.c.) before the surgery. Pups were injected under isoflurane anaesthesia. Mice were placed in a stereotaxic frame, the skin was disinfected with ethanol 70% and betadine, an incision was made, and 1% lidocaine was topically applied to the wound edges for additional local anaesthesia. The bregma and lambda were horizontally aligned. A burr hole was made using a pneumatic dental drill. Injections were targeted to the caudal part of the POM (coordinates from bregma: RC,  $-2.20 \text{ mm}$ ; ML,  $-1.20 \text{ mm}$ ; DV,  $-3.00 \text{ mm}$  for adults; RC,  $-1.45 \text{ mm}$ ; ML,  $-1.60 \text{ mm}$ ; DV,  $-2.90 \text{ mm}$  for pups)<sup>27</sup>. AAV2/1-CAG-ChR2-Venus (200–500 nl for adults;  $50 \text{ nl}$  for pups;  $1.5 \times 10^{12} \text{ GC, UNC Vector Core}$ ; based on the pACAGW-ChR2-Venus-AAV plasmid, Svoboda laboratory, Janelia Farm Research Center)<sup>27,34</sup> or 250–400 nl of AAV2/1-hSynap-GCaMP5G-WPRE-SV40 ( $2.13 \times 10^{13} \text{ GC, Penn Vector Core}$  and the GENIE project)<sup>35</sup> were injected at a maximum rate of  $100 \text{ nl min}^{-1}$  using a glass pipette (Wiretrol, Drummond) attached to an oil hydraulic manipulator (MMO-220A, Narishige). The solution was allowed to diffuse for at least 10 min before the pipette was withdrawn. The craniotomy was filled with Kwik-Cast (WPI) and the skin was re-attached with stainless steel staples (Precise DS15, 3M) or, in the case of AAV-GCaMP injections, a glass window was sealed into the craniotomy as previously described<sup>36</sup>. For cortical GCaMP expression, 20 nl of a mixture of AAV2/9-syn.Flex-GCaMP6s-WPRE-SV40 ( $1.35 \times 10^{13} \text{ GC ml}^{-1}$ ; Penn Vector Core and the GENIE project)<sup>37</sup> and AAV2/1-hSyn-Cre-WPRE-hGH ( $1.04 \times 10^{13} \text{ GC ml}^{-1}$ ; Penn Vector Core) (15,000:1) were injected at a maximum rate of  $10 \text{ nl min}^{-1}$  just before sealing the window. The anaesthesia was reversed with an s.c. injection of a mixture containing atipamezole (Alzane,  $2.5 \text{ mg kg}^{-1}$ ), flumazenil (Anexate,  $0.5 \text{ mg kg}^{-1}$ ), and buprenorphine (Temgesic,  $0.1 \text{ mg kg}^{-1}$ ) in sterile NaCl 0.9% (AFB-mix).

**Cannulation and in vivo photostimulation.** At least 2 weeks after the injection of AAV2-ChR2, mice were anaesthetized with MMF-mix. Intrinsic signal optical imaging was performed as above. A 21-gauge cannula (PlasticsOne) with 2.9 mm of exposed tip was stereotactically inserted through a burr hole (RC,  $-2.20 \text{ mm}$ ; ML,  $-1.20 \text{ mm}$ ; DV,  $-3.00 \text{ mm}$  from bregma) and secured in place with dental cement (Jet Repair Acrylic, Lang Dental Manufacturing). The cannula was closed using a screw cap, and anaesthesia was reversed using AFB-mix. Mice were allowed to recover for 1 day before being prepared for *in vivo* patch-clamp recordings.

For the *in vivo* photostimulation of ChR2-expressing POM neurons, a stripped multimode optical fibre (BFL37-200, Thorlabs) fused to an internal guide (PlasticsOne) was inserted into the cannula. The fibre was coupled to a blue DPSS laser (SDL-473-050MFL, Shanghai Dream Lasers Technology), which was triggered by a pulse-stimulator (Master-8, A.M.P.I.). The rise time of a 1-ms laser pulse (300  $\mu\text{s}$ ) was determined with a high-speed Si photodetector (DET10A, Thorlabs) coupled to an oscilloscope. The power output of the pulses was  $\sim 70\%$  of the steady-state power. The steady-state power at the tip of the fibre was measured using a power meter (PM100D, S120C, Thorlabs) and adjusted before every recording session to  $\sim 40 \text{ mW mm}^{-2}$ . No significant reduction in power was observed at the end of the experiments.

**Ca<sup>2+</sup> imaging and image analysis.** A custom-made stainless steel post was cemented to the skull with dental acrylic (Jet Repair Acrylic, Lang Dental Manufacturing). Imaging was performed 14 to 30 days after virus injection using a custom-built 2-photon laser scanning microscope (<https://openwiki.janelia.org/wiki/display/shareddesigns/Shared+Two-photon+Microscope+Designs>), and using the custom-developed acquisition and microscope control software package ScanImage<sup>38</sup> (<https://openwiki.janelia.org/wiki/display/ephys/ScanImage>)<sup>36</sup>. For imaging of awake mice, mice were trained and habituated to the microscope setup for 7 days before the experiment. During imaging the mice were monitored using an infrared sensitive camera. The GCaMPs were excited using a Ti:sapphire laser (Coherent) tuned to  $\lambda = 910$  nm. For detection we used GaAsP photomultiplier tubes (10770PB-40, Hamamatsu) and a  $\times 20$  (0.8 NA) microscope objective (Olympus). For imaging of axons the field of view typically spanned  $300 \times 300 \mu\text{m}$  (256 lines, 1 ms per line); for dendrites  $43 \mu\text{m} \times 21 - 43 \mu\text{m}$  (64 lines, 0.5 ms per line). The average excitation power was kept below 40 mW, as measured at the focal point of the objective. Bleaching of GCaMPs was negligible. Episodes in which piezo-mediated whisker deflections were immediately followed by active whisking were excluded from the analysis.

All image analyses were performed using custom routines in Matlab. We used cross-correlation based on rigid body translation to register images over time. Small regions of interest (ROIs) were drawn in the dendritic shafts ( $\sim 1 \mu\text{m}^2$ ) or around axonal boutons ( $\sim 7 \mu\text{m}^2$ ), based on averaged and standard deviation images. For each ROI the baseline fluorescence ( $F_0$ ) was calculated based on the mean fluorescence intensity within the selected ROI, averaged over 100 consecutive frames before whisker stimulation. Change in fluorescence ( $\Delta F_i/F_0$ ) was defined as  $(F_i - F_0)/F_0$ , where  $F_i$  is the fluorescence intensity at time  $t$  ( $t = \text{time of the first pixel in each frame}$ ). Boutons were imaged at a depth of  $18 \mu\text{m}$  to  $42 \mu\text{m}$  below the pia, dendrites between  $18 \mu\text{m}$  and  $70 \mu\text{m}$ . The onset of the response was defined as the time in between the whisker stimulus trigger and the time point at which fluorescence intensity reached a  $2 \times$  baseline standard deviation threshold ( $F_0 + (2 \times \text{s.d.})$ ). For extracting spatial and temporal properties of dendritic Ca<sup>2+</sup> events (Extended Data Fig. 5) a Gaussian function was fitted to the fluorescence intensities of the ROIs in a visually 'active' region. All Gaussian fits were normalized to their maximum value. The time course of the change in fluorescence was extracted from the ROI that represented the peak of the Gaussian. A region was considered to be responsive if the fluorescence intensity remained above the threshold ( $F_0 + (2 \times \text{s.d.})$ ) for at least three imaging frames.

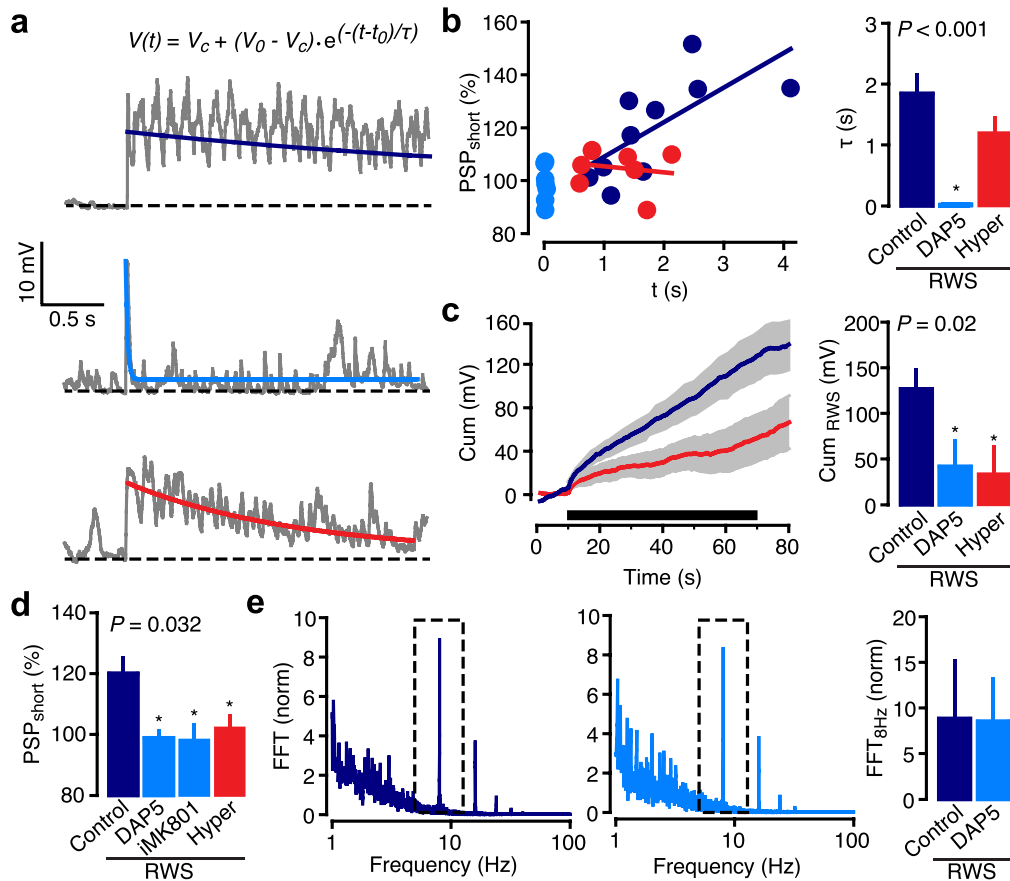
To evaluate the ChR2-Venus expression profiles and the spread of fluorescent mucinol (Extended Data Figs 8 and 9), wide-field epifluorescence images were taken of fixed brain slices. Illumination was set such that the full dynamic range of the 16-bit images was used. A threshold was applied using Fiji's<sup>39</sup> implementation of the Kapur-Sahoo-Wong (Maximum Entropy) method<sup>40</sup>. The resulting image masks were registered to the corresponding coronal plates (ranging from  $-1.94$  to  $-2.70$  mm) of the Paxinos mouse brain atlas<sup>41</sup> using Photoshop (Adobe), at various distances posterior to bregma.

**In vitro whole-cell recordings.** Coronal slices (thickness =  $350 \mu\text{m}$ ) were cut with a vibratome (Leica VT S1000) in ice-cold cutting solution containing (in mM): 83 NaCl, 2.5 KCl, 0.5 CaCl<sub>2</sub>, 3.3 MgSO<sub>4</sub>, 26.2 NaHCO<sub>3</sub>, 1 NaH<sub>2</sub>PO<sub>4</sub>, 22 D-glucose and 72 sucrose. Slices were transferred in normal ACSF at  $\sim 34^\circ\text{C}$  for about 30 min and stored at room temperature before the experiment. ACSF contained (in mM): 124

NaCl, 3 KCl, 2 CaCl<sub>2</sub>, 1.3 MgSO<sub>4</sub>, 26 NaHCO<sub>3</sub>, 1.25 NaH<sub>2</sub>PO<sub>4</sub>, 10 D-glucose with osmolality of 300 mOsm and pH 7.3 when bubbled with 95% O<sub>2</sub> + 5% CO<sub>2</sub>. Individual slices were transferred to the recording chamber and perfused with oxygenated ACSF. All recordings were performed at  $37^\circ\text{C}$  ( $\pm 0.5^\circ\text{C}$ ). Whole-cell recordings were performed using an IR-DIC microscope (Olympus BX51). Recordings were performed using borosilicate glass pipettes with resistance of 4–8 M $\Omega$  and filled with an intracellular solution containing (in mM): 110 K-gluconate, 10 KCl, 10 HEPES, 4 ATP, 0.3 GTP, 10 phosphocreatine and 0.4% biocytin. Recordings were amplified using Multiclamp 700 A amplifiers (Molecular devices, USA), filtered at 4 KHz, digitized (5–20 KHz), and acquired using PulseQ electrophysiology package running on Igor Pro (Wavemetrics, USA). Data processing and analysis was done using Igor (Wavemetrics) and Excel (Microsoft Office). For optogenetic experiments, axons terminals were stimulated with a LED (Thorlabs, Germany) that was focused around the recording electrode using a 4x microscope objective. Drugs used include 4-AP (100  $\mu\text{M}$ ; Sigma Aldrich) and tetrodotoxin (TTX) (1  $\mu\text{M}$ ; Latoxan). To confirm the identity of recorded neurons, 1 mM Alexa 568 hydrazide (Invitrogen) was added to the intracellular solution.

**Statistical analysis.** All statistics were performed using Matlab. The  $\alpha$  significant level was set at 0.05. Normality of all value distributions was assessed by Shapiro–Wilk test ( $\alpha = 0.05$ ). Equality of variance between different distributions was assessed by the Levene median test ( $\alpha = 0.05$ ). Standard parametric tests were only used when data passed the normality and equal variance tests ( $P > 0.05$ ). Non-parametric tests were used otherwise. Only two-sided tests were used. Randomization and blinding methods were not used. No statistical methods were used to estimate sample size, but  $\beta$ -power values were calculated and are provided in Supplementary Information, or in the Extended Data Figure legends.

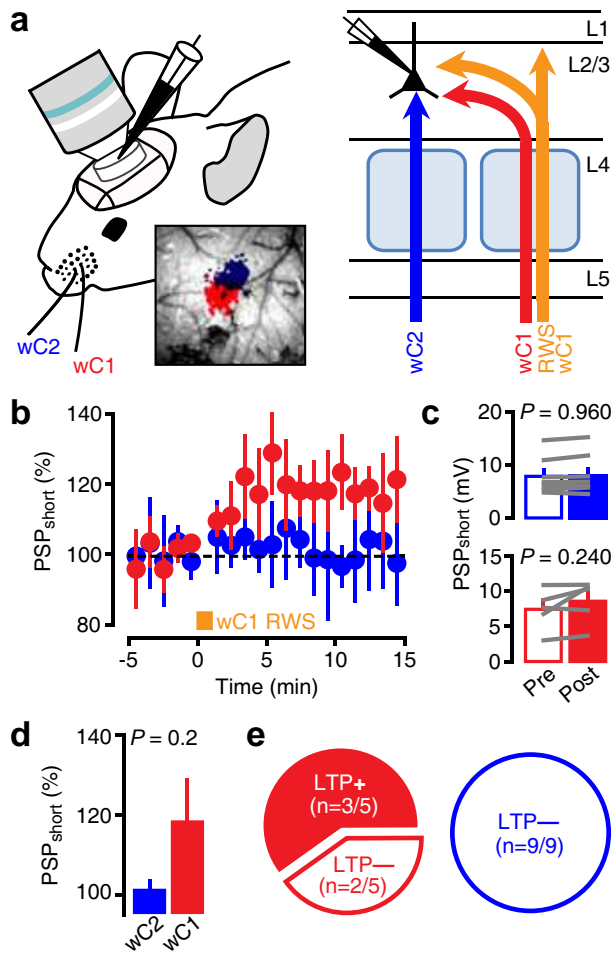
31. Kitamura, K., Judkewitz, B., Kano, M., Denk, W. & Häusser, M. Targeted patch-clamp recordings and single-cell electroporation of unlabeled neurons *in vivo*. *Nature Methods* **5**, 61–67 (2008).
32. Kole, M. H. & Stuart, G. J. Is action potential threshold lowest in the axon? *Nature Neurosci.* **11**, 1253–1255 (2008).
33. Ohno, S. *et al.* A morphological analysis of thalamocortical axon fibers of rat posterior thalamic nuclei: a single neuron tracing study with viral vectors. *Cereb. Cortex* **22**, 2840–2857 (2012).
34. Zhang, F., Wang, L. P., Boyden, E. S. & Deisseroth, K. Channelrhodopsin-2 and optical control of excitable cells. *Nature Methods* **3**, 785–792 (2006).
35. Akerboom, J. *et al.* Optimization of a GCaMP calcium indicator for neural activity imaging. *J. Neurosci.* **32**, 13819–13840 (2012).
36. Holtmaat, A. *et al.* Long-term, high-resolution imaging in the mouse neocortex through a chronic cranial window. *Nature protocols* **4**, 1128–1144 (2009).
37. Chen, T. W. *et al.* Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature* **499**, 295–300 (2013).
38. Polgruto, T. A., Sabatini, B. L. & Svoboda, K. ScanImage: Flexible software for operating laser-scanning microscopes. *Biomed. Eng. Online* **2**, 13 (2003).
39. Schindelin, J. *et al.* Fiji: an open-source platform for biological-image analysis. *Nature Methods* **9**, 676–682 (2012).
40. Kapur, J. N., Sahoo, P. K. & Wong, A. K. C. A new method for gray-level picture thresholding using the entropy of the histogram. *Computer Vis. Graph. Image Process.* **29**, 273–285 (1985).
41. Paxinos, G. & Franklin, K. B. J. *The Mouse Brain in Stereotaxic Coordinates*. 2nd edn (Academic Press, 2001).



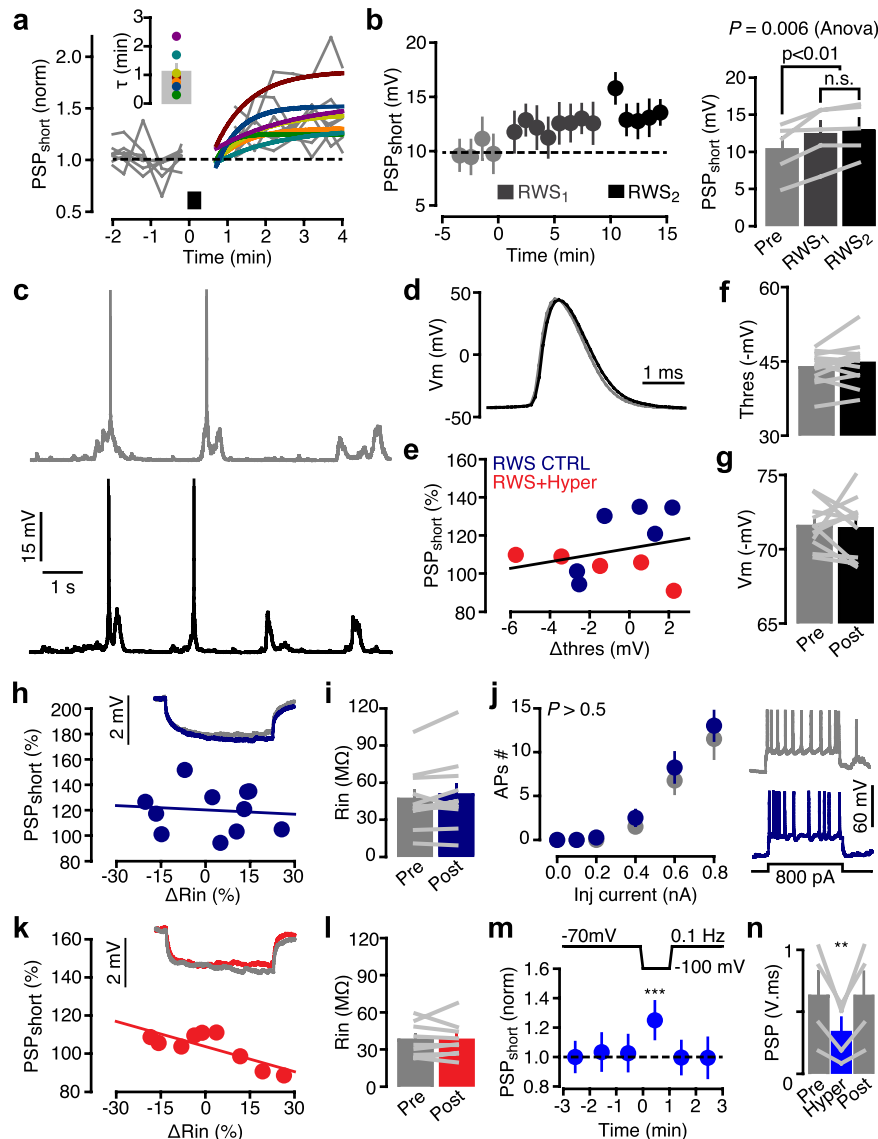
**Extended Data Figure 1 | An NMDAR block suppresses RWS-induced sustained depolarization and prevents LTP.** **a**, Examples of the postsynaptic depolarization as induced by RWS in controls (top), following a blockade of NMDARs by DAP5 (middle), and hyperpolarization (bottom). Only the first 2.5 s (20 deflections) of the recordings are shown. Responses were fit with an exponential, in which  $V(t)$  is the depolarization at time point  $t$  in seconds,  $V_c$  is the depolarization constant (for example, reached after  $>10$  s RWS),  $V_0$  and  $t_0$  are the depolarization and time at RWS onset, and  $\tau$  is the time constant. **b**, Left, under control conditions (dark blue), the level of LTP is linearly correlated to the time constant ( $\tau$ ) of the exponential decay ( $R^2 = 0.49$ ,  $P < 0.05$ ). Following a suppression of NMDAR conductances the time constant and the percentage of LTP are independent (DAP5, light blue,  $R^2 = 0.03$ ,  $P > 0.05$ ; Hyper, red,  $R^2 = 0.03$ ,  $P > 0.05$ ). Each circle represents a single cell. Right, DAP5 significantly reduced the time constant ( $\tau$ ) of the exponential decay ( $P < 0.001$ ; Kruskal–Wallis one-way ANOVA on ranks.  $*P < 0.05$ , post-hoc Dunn’s comparisons versus control condition). **c**, Left, the sustained depolarization during RWS is altered when NMDAR conductances are suppressed by hyperpolarization (red). Black bar indicates the RWS period.

Right, cumulative depolarization at the end of the RWS period (control,  $127 \pm 21$  mV,  $n = 11$ ; +DAP5,  $41 \pm 28$  mV,  $n = 7$ ; Hyper,  $34 \pm 25$  mV,  $n = 9$ ;  $P = 0.02$ , one-way ANOVA ( $\beta = 0.62$ ) and  $*P < 0.05$ , post-hoc Holm–Sidak comparisons versus control condition). **d**, RWS failed to induce LTP when NMDARs are blocked by extracellular application of DAP5, when MK801 is included in the patch pipette, and when cells are hyperpolarized (control,  $119.8 \pm 6\%$ ,  $n = 11$ ; DAP5,  $98.7 \pm 2.5\%$ ,  $n = 7$ ; iMK801,  $98 \pm 5\%$ ,  $n = 3$ ; Hyper,  $103.6 \pm 6\%$ ,  $n = 9$ ;  $P = 0.032$ , one-way ANOVA and  $*P < 0.05$ , post-hoc Holm–Sidak comparisons versus control condition). **e**, A fast Fourier transform (FFT) of the responses during RWS (1 min), in controls (left) and after DAP5 application (middle). The FFT is normalized to the average FFT between 0.1 and 1 Hz. The presence of a strong 8 Hz component after DAP5 indicates that RWS-mediated inputs remain to be activated after the NMDAR block. Right, the magnitude of normalized FFT at 8 Hz is similar between control +DAP5 conditions (control,  $8.8 \pm 8.5$ ,  $n = 11$ ; +DAP5,  $6.3 \pm 4.6$ ,  $n = 7$ ;  $P = 0.733$ , Mann–Whitney  $U$ -test) confirming that part of the whisker deflection-evoked synaptic input was unaffected by DAP5, and follows the rhythmic stimulation. Values in **b–e** are represented as the mean  $\pm$  s.e.m.



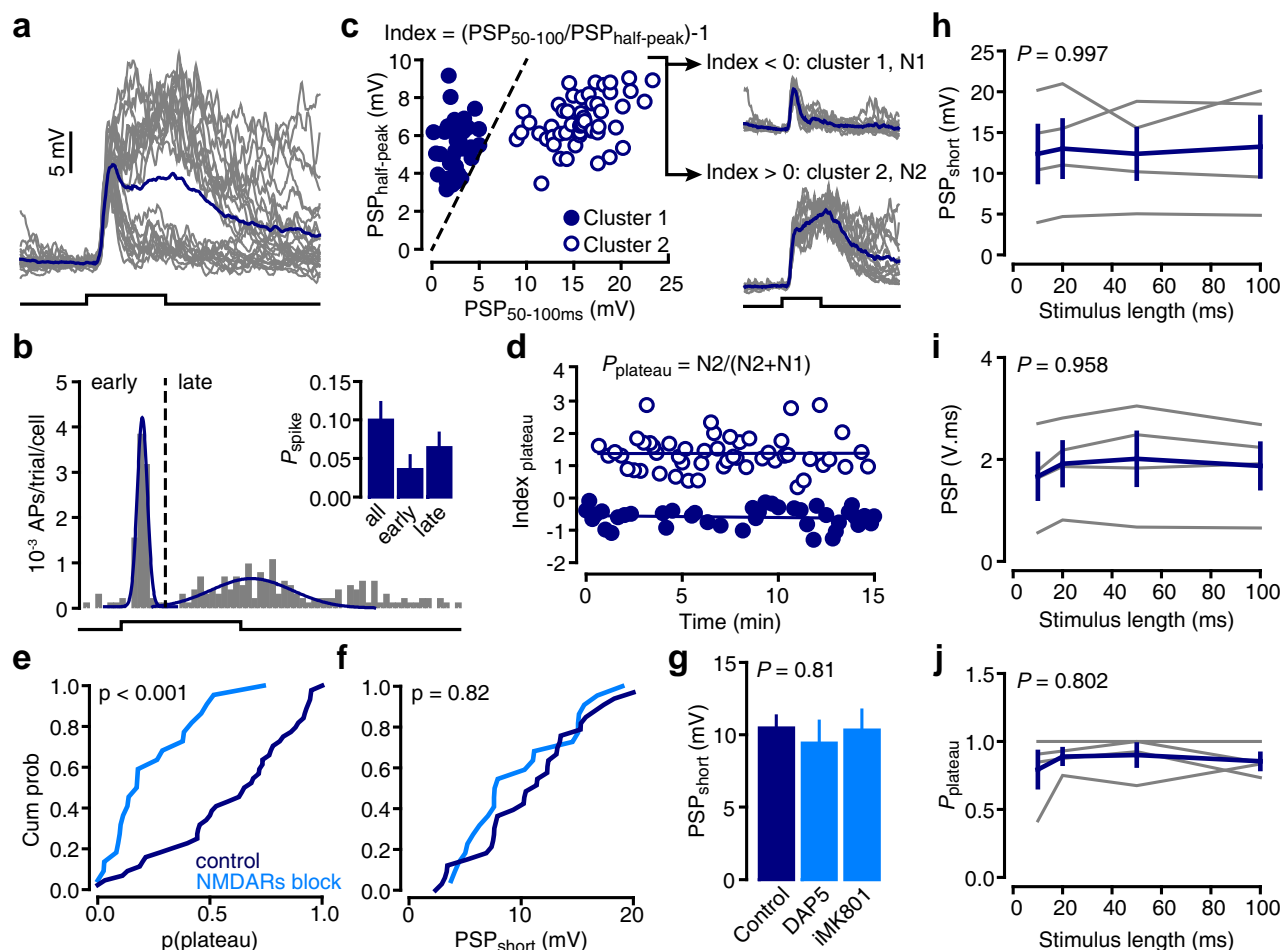


**Extended Data Figure 2 | RWS-induced LTP is column- and whisker-specific.** **a**, Left, schematic of the experiment: whole cell recordings are targeted to the C2 barrel column. Responses are recorded upon deflection of the C2 (principal whisker, wC2) or C1 (surrounding whisker, wC1) whisker. Inset, C2 (blue) and C1 (red) barrel-related columns were mapped using intrinsic optical imaging. Right, schematic of PW and SW-associated synaptic pathways projecting to L2/3 pyramidal cells. After RWS of wC1 (orange), single-whisker deflection PSPs were evoked either by the same whisker that was used for RWS (wC1, red) or by the neighbouring whisker (wC2, blue). **b**, Time course of mean wC2- (blue) and wC1-evoked (red)  $PSP_{short}$  amplitudes ( $\pm$  s.e.m.) following RWS of wC1 (orange bar). **c**, Mean  $PSP_{short}$  amplitude ( $\pm$  s.e.m.) before and after RWS. Top, wC1-RWS did not significantly enhance wC2-evoked mean  $PSP_{short}$  amplitudes (Pre,  $7.8 \pm 1.1$  mV; Post,  $7.9 \pm 1.3$  mV;  $n = 8$ ;  $P = 0.960$ , paired  $t$ -test ( $\beta = 0.05$ )). Bottom, wC1-evoked amplitudes were enhanced in some cells, but despite this positive trend the average difference was not significant (Pre,  $7.4 \pm 1.3$  mV; Post,  $8.6 \pm 1.3$  mV;  $n = 5$ ;  $P = 0.240$ , paired  $t$ -test ( $\beta = 0.15$ )). Grey lines indicate pairs. **d**, **e**, Although the average  $PSP$  amplitude as evoked by wC1 was not significantly different from wC2 (wC2,  $101 \pm 2.5\%$ ,  $n = 8$ ; wC1,  $118 \pm 10\%$ ,  $n = 5$ ;  $P = 0.2$ , Mann-Whitney  $U$ -test; **d**), the number of significantly potentiated cells was higher for wC1 (3 out of 5) than for wC2 (0 out of 9; **e**).



**Extended Data Figure 3 | Characterization of RWS-induced LTP and stability of cell membrane properties.** **a**, For each potentiated cell, a single exponential was fit to the normalized  $PSP_{short}$  amplitudes immediately following RWS, using the following equation:  $PSP(t) = (1 + \Delta PSP_{LTP}) - (\Delta PSP_{LTP} \cdot e^{-(t/\tau)})$ , in which  $PSP(t)$  is the normalized PSP amplitude at time  $t$  in minutes,  $\Delta PSP_{LTP}$  is the average change in PSP amplitude during the LTP phase, and  $\tau$  the time constant (mean  $\tau = 1.09 \pm 0.26$  min; range 0.3–2.34 min;  $n = 7$ ). **b**, Left, time course of mean  $PSP_{short}$  amplitudes following two consecutive RWS protocols (RWS1 and RWS2). Right, the mean  $PSP_{short}$  amplitude increases upon RWS1 but does not further increase upon RWS2, indicating that RWS-evoked LTP is occluded (Pre,  $10.32 \pm 1.6$  mV; RWS1,  $12.35 \pm 1.6$  mV; RWS2,  $12.98 \pm 1.5$  mV;  $n = 5$ ;  $P = 0.006$ , repeated measures ANOVA ( $\beta = 0.97$ ); post-hoc Holm-Sidak comparisons). The error bars represent s.e.m. **c**, Example of typical single-cell membrane potential fluctuations during anaesthesia before (top, grey) and after (bottom, black) RWS. Spontaneous action potentials (APs) are rare and visible only during up states. **d**, Example of spontaneous APs before (grey) and after (black) RWS. APs were time-aligned to their respective threshold. **e**, Effect of RWS on AP thresholds, in cells that displayed some spontaneous APs before and after RWS (but not during RWS). AP threshold was computed as the minimal membrane potential value at the time corresponding to the peak of the third derivative of the membrane potential. The  $\Delta$ threshold was calculated as the difference between the mean AP threshold after RWS and the mean AP threshold before RWS (each circle represents a cell). The level of LTP is independent of  $\Delta$ threshold ( $R^2 = 0.07$ ,  $P > 0.05$ , all cells pooled). **f**, The mean threshold for spontaneous AP is not affected by RWS (Pre,  $-43.7 \pm 0.9$  mV; Post,  $-44.6 \pm 1.2$  mV;  $n = 12$ ;  $P = 0.2$ , paired  $t$ -test ( $\beta = 0.13$ )). **g**, RWS does not

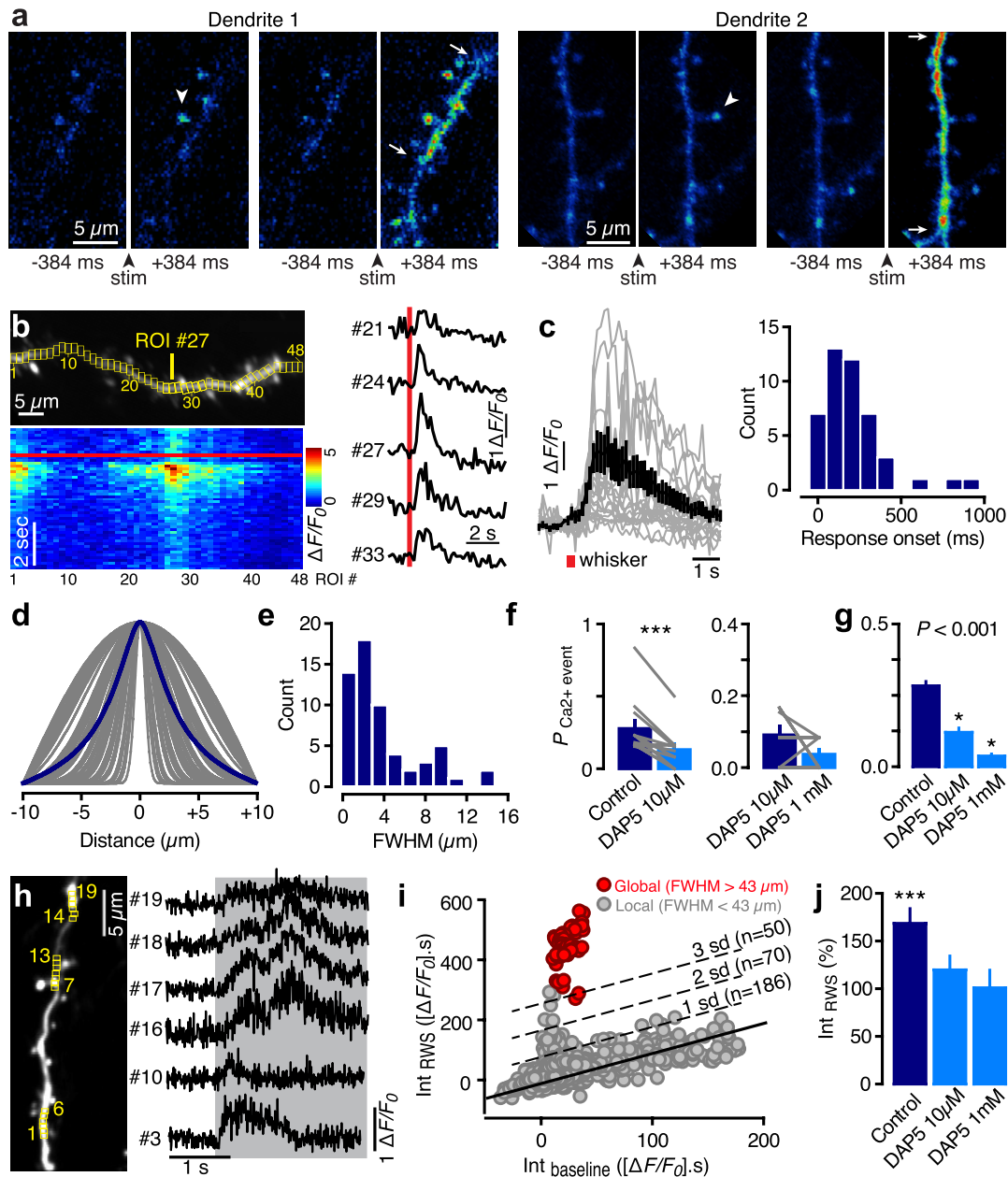
affect the resting membrane potential measured at  $I = 0$  during down states (Pre,  $-71.5 \pm 0.4$  mV; Post,  $-71.4 \pm 0.6$  mV;  $n = 12$ ;  $P = 0.8$ , paired  $t$ -test ( $\beta = 0.05$ )). **h**, The difference between input resistance ( $\Delta Rin$ ) before and after RWS.  $\Delta Rin$  is independent of the level of LTP ( $R^2 = 10^{-3}$ ,  $P > 0.05$ ; each circle represents a cell). Inset, Rin before (grey) and after (dark blue) RWS are estimated by measuring the steady-state resistance of a hyperpolarizing current pulse. **i**, The mean Rin is not affected by RWS (Pre,  $46.6 \pm 7$  M $\Omega$ ; Post,  $50 \pm 8$  M $\Omega$ ;  $n = 11$ ;  $P = 0.13$ , paired  $t$ -test ( $\beta = 0.22$ )). **j**, The number of evoked APs as a function of injected somatic current injection is not significantly modified by RWS (two-way ANOVA,  $P > 0.5$ ,  $n = 4$ ). **k**, The relationship between  $\Delta Rin$  and the relative change in amplitude of  $PSP_{short}$ . Each circle represents a cell that was hyperpolarized only during RWS (RWS + Hyper,  $R^2 = 0.6$ ). **l**, The mean Rin is not affected by RWS + Hyper (Pre,  $37.4 \pm 4$  M $\Omega$ ; Post,  $37.7 \pm 5$  M $\Omega$ ;  $n = 9$ ;  $P = 0.9$ , paired  $t$ -test ( $\beta = 0.05$ )). **m**, **n**, The  $PSP_{short}$  amplitude evoked by low frequency (0.1 Hz) single whisker deflections, before (Pre), during (Hyper) and after (Post) hyperpolarization ( $-100$  mV). Hyperpolarization increases the PSP amplitude (**m**) due to an enhanced driving force (Pre,  $11.3 \pm 2.5$  mV (average  $-3$  to  $-1$  min); Hyper,  $14.7 \pm 3$  mV; Post (average  $+1$  to  $+2$  min),  $11.3 \pm 2.4$  mV;  $n = 4$ ;  $P = 0.002$ , one-way repeated-measures ANOVA ( $\beta = 0.98$ );  $***P < 0.001$ , Holm-Sidak post-hoc comparisons Hyper versus Pre and Post conditions). However, the integrated PSP is significantly reduced due to the absence of plateau potentials (**n**) (Pre,  $0.625 \pm 0.2$  V $\cdot$ ms; Hyper,  $0.335 \pm 0.1$  V $\cdot$ ms; Post,  $0.626 \pm 0.2$  V $\cdot$ ms;  $n = 4$ ;  $P = 0.011$ , one-way repeated-measures ANOVA ( $\beta = 0.86$ );  $**P = 0.007$ , Holm-Sidak post-hoc comparisons Hyper versus Pre and Post conditions).



**Extended Data Figure 4 | The extraction of whisker-evoked plateau potentials.** **a**, Individual PSPs (grey lines) and the average PSP (dark blue line) in a single cell in response to single principal whisker deflections (100-ms deflections). Individual traces show short and long-latency components. The responses to the 30 successive deflections reveal two whisker-evoked PSP populations: one that only contains short-latency PSPs and a second population that contains both short and long-latency PSPs. **b**, Similar to non-spiking cells, the distribution of whisker-evoked action potentials (APs) also reveals two populations of spikes, based on their onset delay. Inset, for each spiking cell, whisker-evoked spikes were sorted as early and late spikes, according to the delay of the first peak of the subthreshold response. The corresponding probabilities were then computed (early,  $P = 0.04 \pm 0.02$ ,  $n = 15$ ; late,  $P = 0.06 \pm 0.02$ ,  $n = 15$ ;  $P = 0.345$ ,  $z$ -test). The equal probabilities indicate that L2/3 cells spike as often upon a long-latency depolarization as in response to a short-latency PSP. **c**, Left, for each trial, the relationship between the PSP half-peak amplitude and the average membrane potential between 50 and 100 ms after the onset reveals two distinct clusters. Dotted line represents the identity line. Right, cluster 1 (top) is defined by an index  $< 0$  and consists of PSPs containing only a short latency PSP that quickly returns to

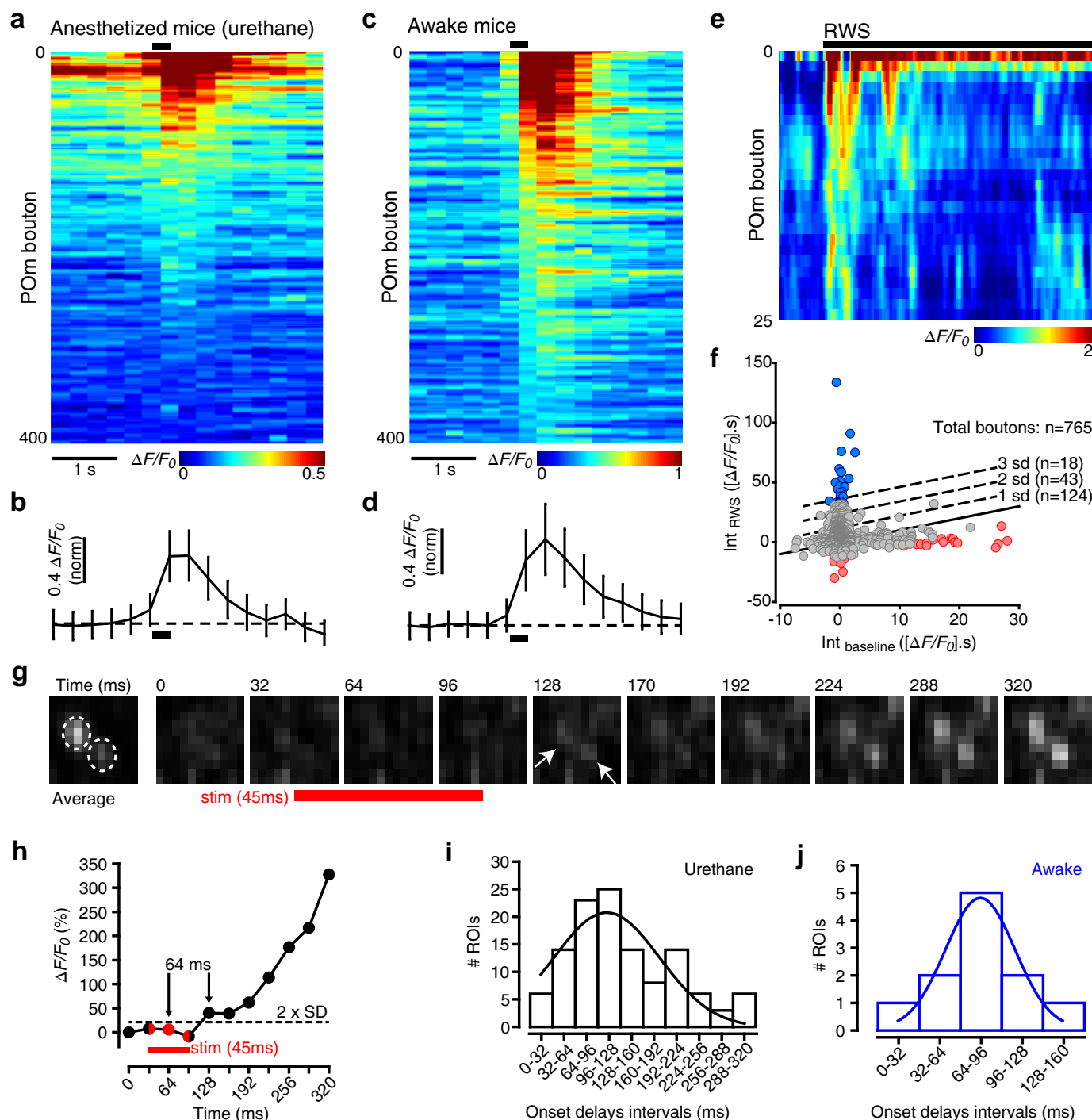
the resting membrane potential. Cluster 2 (bottom) is defined by an index  $> 0$  and consists of compound PSPs with short and long-latency components. The long-latency component of the PSP has a strong plateau-like appearance. Therefore, we defined this index as the  $\text{Index}_{\text{plateau}}$ . **d**, Example of the distribution of clusters 1 and 2 over time. The  $\text{Index}_{\text{plateau}}$  was computed from the example shown in **a**. The probability of eliciting plateau potentials in a neuron is calculated by dividing the number of PSPs in cluster 2 ( $N_2$ ) by the total number of PSPs ( $N_1 + N_2$ ). **e**, **f**, Blocking NMDARs (light blue) significantly reduced the probability of eliciting plateau potentials as compared to controls (dark blue) (**e**, ks test,  $P < 0.001$ ), but does not affect the amplitude of the short-latency PSP ( $\text{PSP}_{\text{short}}$ ) (**f**, ks test,  $P > 0.05$ ). **g**, Blocking NMDARs by epidural application of DAP5 or intracellular MK801 does not affect the amplitude of  $\text{PSP}_{\text{short}}$  (control,  $10.5 \pm 0.8$  mV,  $n = 33$ ; DAP5,  $9.5 \pm 1.5$  mV,  $n = 12$ ; iMK801,  $10.3 \pm 1.3$  mV,  $n = 10$ ;  $P > 0.05$ , one-way ANOVA ( $\beta = 0.05$ )). **h**–**j**, Changes in the length of the whisker deflection period (stimulus length) do not affect the whisker-evoked mean  $\text{PSP}_{\text{short}}$  amplitude (**h**), the mean PSP integral (**i**), or the probability to elicit plateau potentials (**j**) ( $n = 4$ ,  $P > 0.05$ , one-way ANOVA ( $\beta = 0.05$ )). Values in **b** and **g**–**j** are represented as the mean  $\pm$  s.e.m.





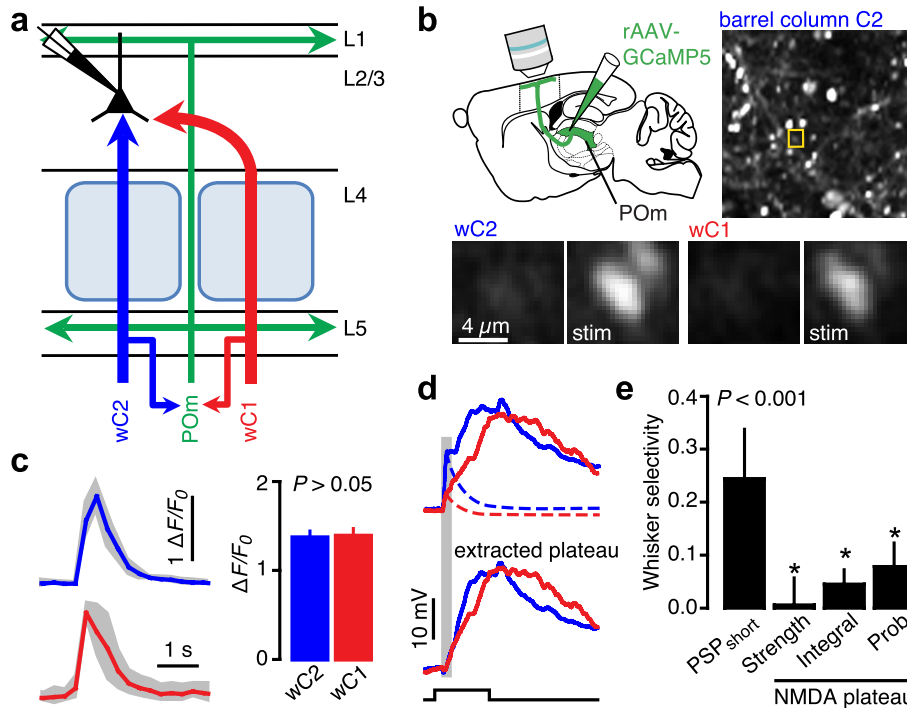
**Extended Data Figure 5 | NMDAR-dependent whisker-evoked  $\text{Ca}^{2+}$  events in L2/3 pyramidal neuron dendritic tufts.** **a**, Examples of single whisker deflection-evoked  $\text{Ca}^{2+}$  responses in dendritic spines (arrowheads) and dendritic shafts (regions between arrows). **b**, Top left, GCaMP6 s fluorescence standard deviation image with ROIs. Bottom left, raster plot of  $\Delta F/F_0$  for each ROI in the top panel (aligned). Red bar represents the whisker stimulation onset. Right,  $\Delta F/F_0$  traces of some ROIs from **a**. **c**, Left, time course of mean dendritic  $\text{Ca}^{2+}$  response in individual dendritic branches upon a single whisker deflection (2–5 trials per branch,  $n = 48$  dendritic branches,  $n = 9$  mice). Grey, individual branches. Black, average response. Right, distribution of response onset times. **d**, Averaged (blue thick line) and individual (grey lines) Gaussian fits of local responsive regions in dendritic shafts. **e**, Distribution of the FWHM of the Gaussian fits in **d**. **f**, Epidural application of DAP5 (10  $\mu\text{M}$ ) significantly reduces whisker-evoked local dendritic  $\text{Ca}^{2+}$  response probabilities. The subsequent addition of 1 mM DAP5 in some cases further reduced probabilities and in others did not show an additive effect (control,  $0.29 \pm 0.06$ , DAP5 (10  $\mu\text{M}$ ),  $0.14 \pm 0.04$ ;  $n = 16$  branches,  $n = 3$  mice;  $P = 0.002$ , Wilcoxon signed-rank test; DAP5 (10  $\mu\text{M}$ ),

$0.08 \pm 0.03$ ; DAP5 (1 mM),  $0.04 \pm 0.02$ ;  $n = 9$  branches,  $N = 2$  mice;  $P = 0.37$ , Wilcoxon signed-rank test). **g**, On average, DAP5 significantly reduces whisker-evoked local dendritic  $\text{Ca}^{2+}$  response probabilities (control,  $0.28 \pm 0.03$ ; DAP5 (10  $\mu\text{M}$ ),  $0.12 \pm 0.03$ ; DAP5 (1 mM),  $0.04 \pm 0.02$ ; Kruskal–Wallis one-way ANOVA on ranks; post-hoc Dunn’s comparisons versus control condition,  $P < 0.05$ ). **h**, Left, GCaMP6 s fluorescence standard deviation image with ROIs. Right,  $\Delta F/F_0$  traces of some ROIs from the left panel. Grey box represents the RWS period. **i**, Integrated  $\Delta F/F_0$  in dendritic branches during RWS (0–15 s) as a function of the response before RWS (0–15 s baseline). Each circle represents a single dendritic branch. Red, global events (responses spanning the whole field of view, minimally 43  $\mu\text{m}$ ); grey, local events (responses spanning a portion of the field of view, maximally 43  $\mu\text{m}$ ). Black line indicates the identity line. RWS significantly increases  $\Delta F/F_0$  for a substantial number of branches. **j**, The average integrated  $\Delta F/F_0$  in dendritic branches during RWS (0–15 s) is significantly reduced upon topical application of DAP5 (control,  $168 \pm 15\%$ ; DAP5 (10  $\mu\text{M}$ ),  $119.6 \pm 14.5\%$ ; DAP5 (1 mM),  $100.1 \pm 18.5\%$ ; paired  $t$ -test).



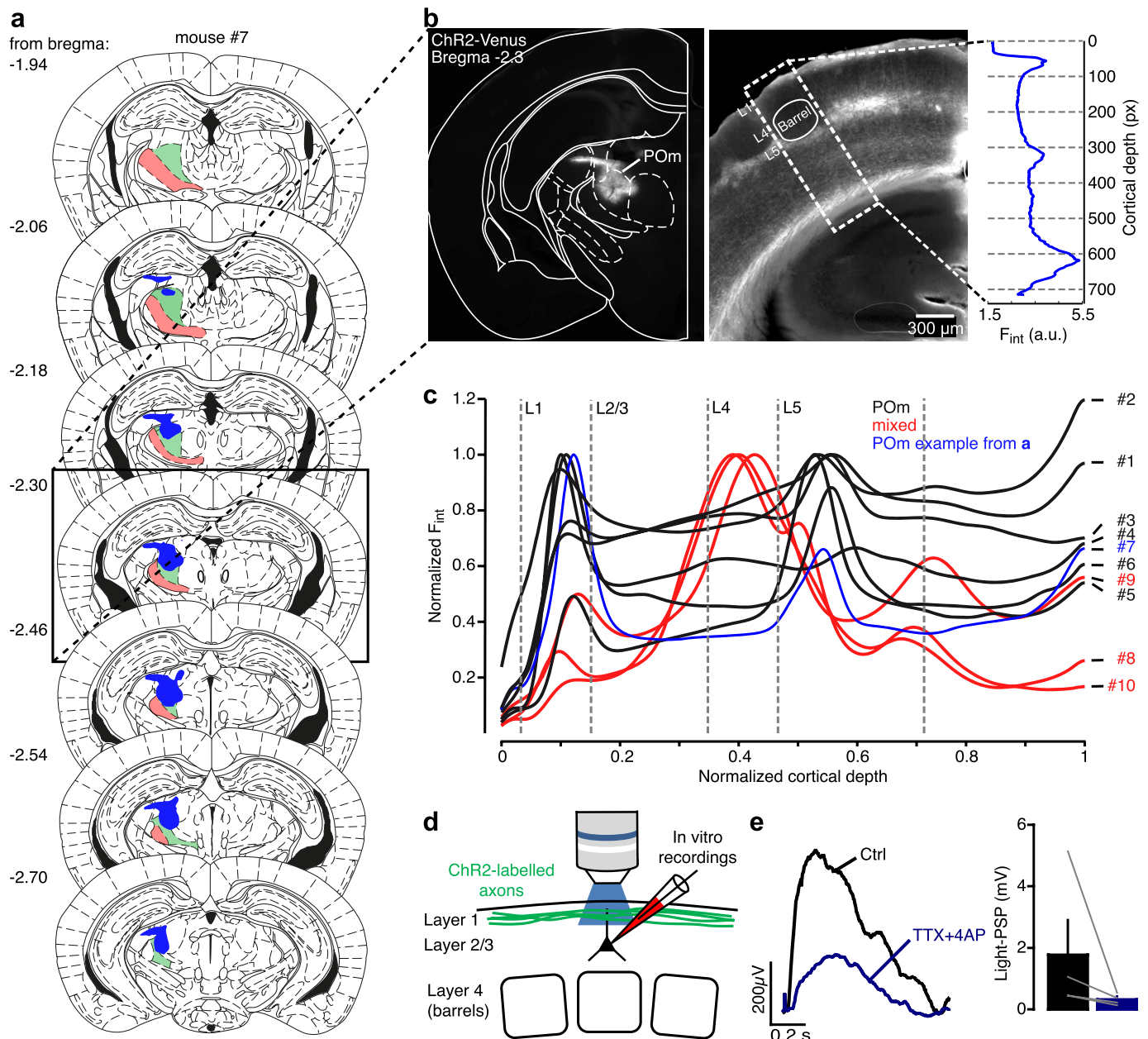
**Extended Data Figure 6 | Whisker-evoked activity in cortical POM efferents in anaesthetized and awake mice.** **a, c**, Raster plots of average  $\text{Ca}^{2+}$  responses ( $\Delta F/F_0$ ) in 400 different cortical POM efferent boutons over 10–20 successive (0.1 Hz) bursts of 5 whisker deflections (20 Hz, black bar) in anesthetized mice (**a**) and in awake mice (**c**). **b, d**, Average whisker-evoked responses over 400 boutons under anaesthesia (**b**) and under wakefulness (**d**). **e**, Raster plot for 25 of the most active boutons (out of 765) upon RWS. Black bars indicate the period of RWS (1 min). The  $\text{Ca}^{2+}$  response in some boutons remains elevated over the whole RWS period. **f**, Integrated  $\text{Ca}^{2+}$  responses of individual boutons during RWS (0–15 s) as a function of their responses before RWS (15 s baseline). Each circle represents a single bouton. The 25 most and least responsive boutons are in blue and red respectively. Black

line indicates the mean relationship (linear fit). Dotted lines indicate the relationship at various standard deviations from the fit. A substantial proportion of boutons (16%; 124 out of 765) display RWS/baseline ratios larger than 1 s.d. from the mean. **g**, Time-lapse image of fluorescence change representing  $\text{Ca}^{2+}$  responses in axonal boutons (dotted circles) upon a single whisker deflection (red bar; 45 ms). Response onsets are indicated by arrows. Scale bar represents 1  $\mu\text{m}$ . **h**, Example of the response curve of the boutons in **g**. Response onset latency was defined as the time frame in which  $\Delta F/F_0$  exceeded  $2 \times \text{s.d.}$  of the baseline. **i, j**, Distributions of response onset latencies under anaesthesia (black, 120 trials,  $n=5$  boutons,  $n=3$  mice) and under wakefulness (blue, 11 trials,  $n=5$  boutons,  $n=3$  mice).



**Extended Data Figure 7 | Plateau potentials in L2/3 pyramidal neurons and POM-efferent activity are not whisker-specific.** **a**, Schematic of the experiment. Whole-cell recordings are targeted to the C2 barrel column. Responses are recorded upon deflection of the C2 whisker (wC2, principal whisker, PW, blue) or the C1 whisker (wC1, surrounding whisker, SW, red). **b**, Right, example of 2PLSM images of POM boutons expressing GCaMP5. Both the PW (wC2) and SW (wC1) evoke a  $\text{Ca}^{2+}$  response. **c**, Left, the average  $\text{Ca}^{2+}$  transient ( $\Delta F/F_0$ ) for both whisker deflections (shadows represent the s.d.). Right, mean  $\Delta F/F_0$  upon deflection of the PW (wC2) and SW (wC1) (wC2,  $1.37 \pm 0.07$ ,  $n = 5$ ; wC1,  $1.38 \pm 0.08$ ,  $n = 5$ ;  $P > 0.05$ ). This confirms that POM activity is not selective for whiskers. Values are represented as mean  $\pm$  s.e.m. **d**, Top, example of the average PSP evoked by the PW (blue) or the SW (red). To estimate the integral of the plateau potential (bottom), the decay of the first component is fitted with a single exponential and subtracted from the

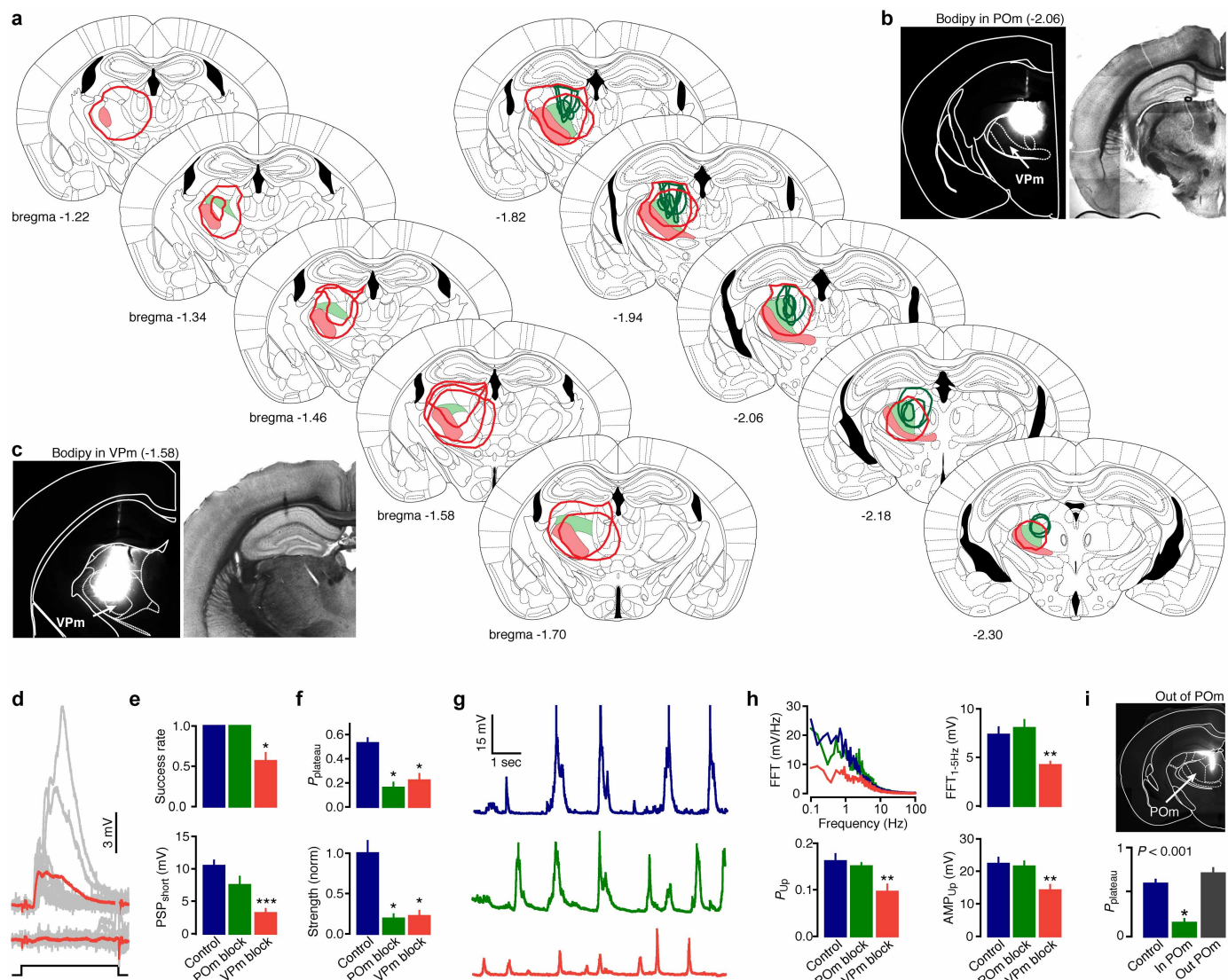
average of PSPs containing both short and late-latency components (cluster 2 in Extended Data Fig. 4c). **e**, For each parameter, whisker selectivity is defined by the ratio between the PW and SW:  $(\text{PW} - \text{SW})/(\text{PW} + \text{SW})$ . All parameters related to plateau potentials (plateau strength, plateau integral, probability) are not specific to either one of the whiskers. In contrast, and as expected, the short-latency PSP amplitude ( $\text{PSP}_{\text{short}}$ ) is higher (and thus more selective) for the PW ( $\text{PSP}_{\text{short}}$ ,  $0.25 \pm 0.09$ ,  $n = 27$ ; plateau strength,  $0.006 \pm 0.05$ ,  $n = 26$ ; plateau integral,  $0.04 \pm 0.02$ ,  $n = 26$ ; plateau probability,  $0.07 \pm 0.04$ ,  $n = 27$ ;  $P < 0.001$ , one-way ANOVA on ranks;  $*P < 0.05$ , post-hoc Dunn's comparisons versus  $\text{PSP}_{\text{short}}$  condition). As the amplitude of short-latency PSPs is whisker-selective and plateau potentials are not, it is conceivable that POM associated synaptic pathways are responsible for mediating whisker-evoked plateau potentials.



**Extended Data Figure 8 | AAV-mediated expression of ChR2-Venus in the POM nuclei of the thalamus and their efferents in L1.** **a**, Representative example of the ChR2-Venus expression profile in the mouse thalamus. The expression profile of ChR2-Venus is depicted in blue, POM nuclei in light green, and the VPm in light red. **b**, Left, example of a cortical slice with ChR2-Venus fluorescence in the caudal sector of the POM (bregma -2.3). Right, an image of the fluorescence profile in the somatosensory cortex in the same animal as on the left. Fluorescence intensities ( $F_{int}$ ) were measured as a function of cortical depth (in pixels (px)) by summing all pixels within the dotted rectangle over the short axis (pixel size = 1.85  $\mu$ m). The fluorescence intensity profile is similar to the cortical projection pattern of efferents from the caudal sector of the POM (Supplementary Note 4). **c**, Plot comparing the intensity profiles of 10 different animals in which injections were aimed at the POM nuclei in the thalamus. Animals classified as bearing expression profiles in the barrel cortex that are typical of POM projections (Supplementary

Notes 4 and 5) displayed distinct fluorescence peaks in L1 (0.04 – 0.12 normalized depth) and L5 (0.48 – 0.7 normalized depth;  $n = 7$ , black lines and blue line). In contrast, animals with a (additional) distinct peak in L4 (0.36 – 0.48 normalized depth;  $n = 3$ , red lines) were considered as having at least some spurious expression in VPm, and were thus excluded from the analysis in Fig. 3. The classification matched the expression profiles in the thalamus (Supplementary Information). **d**, **e**, Assessment of POM mediated synaptic inputs onto L2/3 neurons in acute cortical slice preparations. **e**, Schematic of the slice experiment. L2/3 pyramidal neurons were recorded during local photostimulation (through the objective) of ChR2-expressing POM axons. **f**, Left, example of photostimulation-evoked PSPs in a single L2/3 cell under control conditions and following bath application of TTX and 4AP. Right, average PSP amplitudes in controls and after TTX + 4AP application (control,  $1.8 \pm 1.1$ ; TTX + 4AP,  $0.31 \pm 0.08$ ,  $n = 4$ ,  $P = 0.125$ , paired Wilcoxon signed-rank test).





**Extended Data Figure 9 | The spread of fluorescent muscimol in thalamic nuclei.** **a**, Coronal diagrams of the mouse brain adapted from the Paxinos atlas<sup>41</sup> including the POM (light green) and VPM (light red) nuclei at various posterior distances from bregma. Each red or green line represents the maximal spread of fluorescent muscimol, as assessed using whole field epifluorescence microscopy (Olympus;  $\times 20$  and  $\times 60$  objective) and Neurolucida (MicroBrightfield) reconstructions. Green lines (8 mice) represent injections that were confined to the POM. Red lines (3 mice) represent injections that infiltrated both POM and VPM thalamic nuclei. **b**, **c**, Examples of a muscimol injection in the caudal part of the POM (**b**) and an injection that spread into both POM and VPM nuclei (**c**). **d**, Examples of whisker-evoked PSPs in L2/3 neurons of mice in which muscimol was present in both POM and VPM. Successes (top) and failures (bottom) are shown. Grey lines, individual trials, light red lines, average. **e**, Blocking activity in VPM decreases the whisker-evoked PSP success rate (top; control,  $1 \pm 0$ ,  $n = 33$ ; POM block,  $1 \pm 0$ ,  $n = 9$ ; POM + VPM block,  $0.56 \pm 0.1$ ,  $n = 7$ ;  $P < 0.001$ , one-way ANOVA on ranks;  $*P < 0.05$ , post-hoc Dunn's comparisons versus control condition), as well as PSP<sub>short</sub> amplitudes (bottom; control,  $10.5 \pm 0.8$ ,  $n = 33$ ; POM block,  $7.5 \pm 1.3$ ,  $n = 9$ ; POM + VPM block,  $3.15 \pm 0.6$ ,  $n = 7$ ;  $P < 0.001$ , one-way ANOVA;  $*P < 0.05$ , post-hoc Holm-Sidak's comparisons versus control condition). **f**, Blocking POM + VPM or POM only significantly decreases the probability of plateau potentials (top; control,  $0.35 \pm 0.04$ ,  $n = 33$ ; POM block,  $0.16 \pm 0.04$ ,  $n = 9$ ; VPM block,  $0.22 \pm 0.05$ ,  $n = 7$ ;  $P < 0.001$ , one-way ANOVA on ranks;  $*P < 0.05$ , post-hoc Dunn's comparisons versus control condition), and the normalized plateau strength (bottom; control,  $1 \pm 0.15$ ,  $n = 33$ ; POM block,  $0.19 \pm 0.05$ ,  $n = 9$ ; POM + VPM block,

$0.22 \pm 0.07$ ,  $n = 7$ ;  $P < 0.001$ , one-way ANOVA on ranks;  $*P < 0.05$ , post-hoc Dunn's comparisons versus control condition). **g**, examples of single-cell spontaneous membrane potential fluctuations during anaesthesia in controls (top) and upon muscimol injections in POM (middle) or POM + VPM (bottom). **h**, Top left, FFT of membrane potentials in controls (dark blue) and after muscimol injection in POM (green) or POM + VPM (light red). Top right, Blocking POM + VPM significantly decreases the 1–5-Hz range in the FFT (control,  $7.33 \pm 0.76$ ,  $n = 14$ ; POM block,  $8.03 \pm 0.81$ ,  $n = 8$ ; VPM block,  $4.2 \pm 0.37$ ,  $n = 7$ ;  $P = 0.008$ , one-way ANOVA;  $*P < 0.05$ , post-hoc Holm-Sidak's comparisons versus control condition). Bottom, blocking POM + VPM significantly decreases the probability of spontaneous up states (left ( $P_{up}$ ); control,  $0.16 \pm 0.016$ ,  $n = 14$ ; POM block,  $0.15 \pm 0.008$ ,  $n = 8$ ; POM + VPM block,  $0.09 \pm 0.016$ ,  $n = 7$ ;  $P = 0.023$ , one-way ANOVA;  $*P < 0.05$ , post-hoc Holm-Sidak's comparisons versus control condition), as well as the amplitude of spontaneous up states (right ( $AMP_{up}$ ); control,  $22.3 \pm 1.9$ ,  $n = 14$ ; POM block,  $21.5 \pm 1.6$ ,  $n = 8$ ; POM + VPM block,  $14.1 \pm 1.7$ ,  $n = 7$ ;  $P = 0.02$ , one-way ANOVA;  $*P < 0.05$ , post-hoc Holm-Sidak's comparisons versus control condition). **i**, Animals in which muscimol injections in the medial posterior thalamus did not infiltrate the POM (out of POM) were used as a negative controls ( $n = 6$ ). In these animals, the probability of eliciting plateau potentials ( $P_{plateau}$ ) remained equal to controls. The probability was significantly reduced, only when muscimol was correctly targeted to POM (control,  $0.6 \pm 0.04$ ,  $n = 44$ ; +muscimol In POM,  $0.16 \pm 0.04$ ,  $n = 9$ ; +muscimol Out POM,  $0.72 \pm 0.06$ ,  $n = 8$ ;  $P < 0.001$ , one-way ANOVA on ranks;  $*P < 0.05$ , post-hoc Dunn's comparisons versus control condition). The values represent the mean  $\pm$  s.e.m.

# Luminal signalling links cell communication to tissue architecture during organogenesis

Sevi Durdu<sup>1</sup>, Murat Iskar<sup>1</sup>, Celine Revenu<sup>1†</sup>, Nicole Schieber<sup>1</sup>, Andreas Kunze<sup>1</sup>, Peer Bork<sup>1</sup>, Yannick Schwab<sup>1</sup> & Darren Gilmour<sup>1</sup>

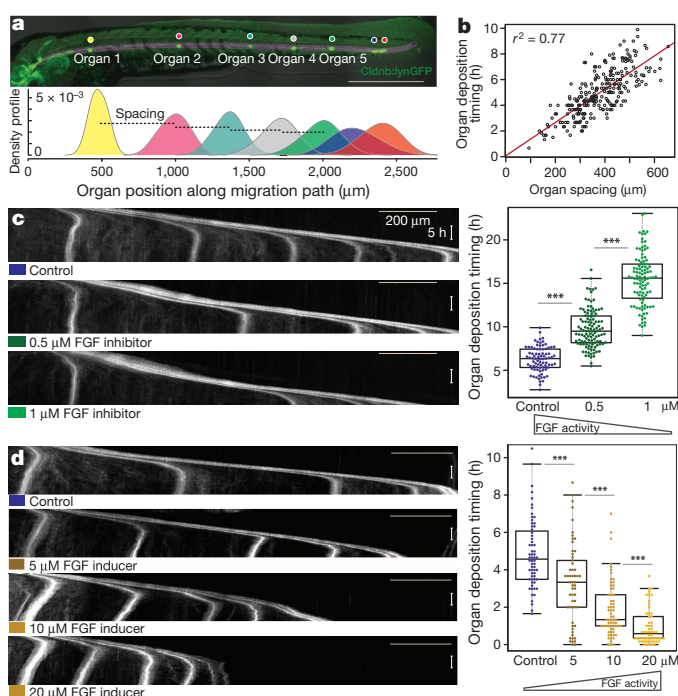
Morphogenesis is the process whereby cell collectives are shaped into differentiated tissues and organs<sup>1</sup>. The self-organizing nature of morphogenesis has been recently demonstrated by studies showing that stem cells in three-dimensional culture can generate complex organoids, such as mini-guts<sup>2</sup>, optic-cups<sup>3</sup> and even mini-brains<sup>4</sup>. To achieve this, cell collectives must regulate the activity of secreted signalling molecules that control cell differentiation, presumably through the self-assembly of microenvironments or niches. However, mechanisms that allow changes in tissue architecture to feedback directly on the activity of extracellular signals have not been described. Here we investigate how the process of tissue assembly controls signalling activity during organogenesis *in vivo*, using the migrating zebrafish lateral line primordium<sup>5</sup>. We show that fibroblast growth factor (FGF) activity within the tissue controls the frequency at which it deposits rosette-like mechanosensory organs. Live imaging reveals that FGF becomes specifically concentrated in microluminal structures that assemble at the centre of these organs and spatially constrain its signalling activity. Genetic inhibition of microlumen assembly and laser micropuncture experiments demonstrate that microlumina increase signalling responses in participating cells, thus allowing FGF to coordinate the migratory behaviour of cell groups at the tissue rear. As the formation of a central lumen is a self-organizing property of many cell types, such as epithelia<sup>6</sup> and embryonic stem cells<sup>7</sup>, luminal signalling provides a potentially general mechanism to locally restrict, coordinate and enhance cell communication within tissues.

A major challenge in biology is to explain how the pattern of complex organs emerges through dynamic self-organizing processes occurring at cellular and molecular scales<sup>1,8,9</sup>. The development of the zebrafish posterior lateral line system provides an example of an *in vivo* organogenesis process that has the potential to be understood quantitatively at subcellular resolution<sup>10</sup>. Here, a series of rosette-like mechanosensory organs is assembled and deposited along the flanks of the embryo by a collectively migrating epithelial primordium<sup>5</sup>. While a number of signalling pathways required for this process have been identified<sup>11</sup>, it is currently not known how their activity is coupled to this organogenesis process. We therefore first performed a quantitative analysis of the normal organ deposition process by time-lapse imaging of many wild-type (WT) embryos (Fig. 1a and Supplementary Videos 1 and 2). This revealed that the overall pattern of organ spacing is determined by the timing of deposition events, rather than by sustained changes in the speed of primordium migration or growth of the embryo (Fig. 1b and Extended Data Fig. 1).

The best candidate regulator of this organ deposition process is FGF signalling, as FGF ligands have been shown to be required for organ formation<sup>12–14</sup>. To test if this pathway controls organ deposition timing we reduced its activity in a stepwise manner, by titrating the FGF receptor inhibitor SU5402 (ref. 12). This showed that reducing FGF activity results in a dose-dependent delay in organ deposition (Fig. 1c, Extended Data Fig. 2 and Supplementary Video 3), a finding we confirmed using mutants for *Fgfr1a*<sup>15</sup>, the receptor that mediates signalling in this context (Extended Data Fig. 2 and Supplementary Video 4). Conversely, when we increased the concentration of FGF-ligand, by expressing a

progesterone-inducible transcription factor<sup>16</sup> (*cxc4b:lexPR*) that drives uniform overexpression of a functional fusion protein of Fgf3 and green fluorescent protein (*lexOP:fgf3-GFP*), organ deposition was accelerated in a dose-dependent manner (Fig. 1d, Extended Data Fig. 2 and Supplementary Video 5). Uniform overexpression of Fgf3-GFP did not significantly alter rosette-like organ assembly rate, indicating that its effect was primarily on the migratory behaviour of assembled organs (Extended Data Fig. 3). Thus, the timing of this organ deposition process can be controlled over a wide dynamic range by the activity level of a single signalling molecule.

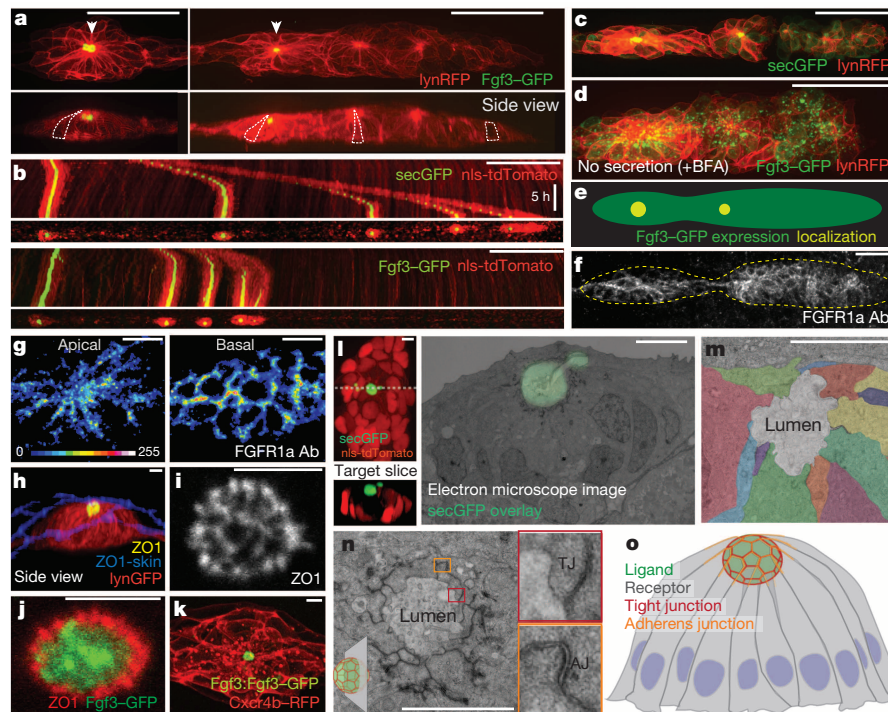
Since FGF regulates lateral line organ deposition in a dose-dependent manner, its extracellular concentration and distribution must be tightly controlled. Imaging the FGF distribution after uniform overexpression



**Figure 1 | FGF signalling regulates organ deposition timing in a dose-dependent manner.** **a**, Quantitative analysis of lateral line patterning. Position of organs along lateral line at 2 days post-fertilization (d.p.f.) (*cldnb:lynGFP*). Plot shows intensity profile of pooled organ positions (below,  $N = 60$  (throughout,  $N$  represents number of embryos and  $n$  represents data points)). Organs and migrating primordium are colour-coded. **b**, Correlation of organ deposition timing and spacing between consecutive depositions (Spearman  $r^2 = 0.77$ ,  $n = 260$ ). **c**, **d**, Influence of FGF level on organ deposition. **c**, Kymographs of control, 0.5 μM and 1 μM SU5402-treated samples. Plot shows quantification of organ deposition timing ( $n = 82, 114, 104$ ). **d**, Kymographs of control, 5, 10 and 20 μM RU486-treated samples and plots of organ deposition timing ( $n = 64, 57, 65, 58$ ). Scale bars, 500 μm (a), 200 μm, 5 h (c, d). Statistics: Wilcoxon, \*\*\* $P < 0.001$ .

<sup>1</sup>European Molecular Biology Laboratory Heidelberg, Meyerhofstrasse 1, 69117 Heidelberg, Germany. <sup>†</sup>Present address: Institut Curie, 26 rue d'Ulm, 75248 Paris, France.





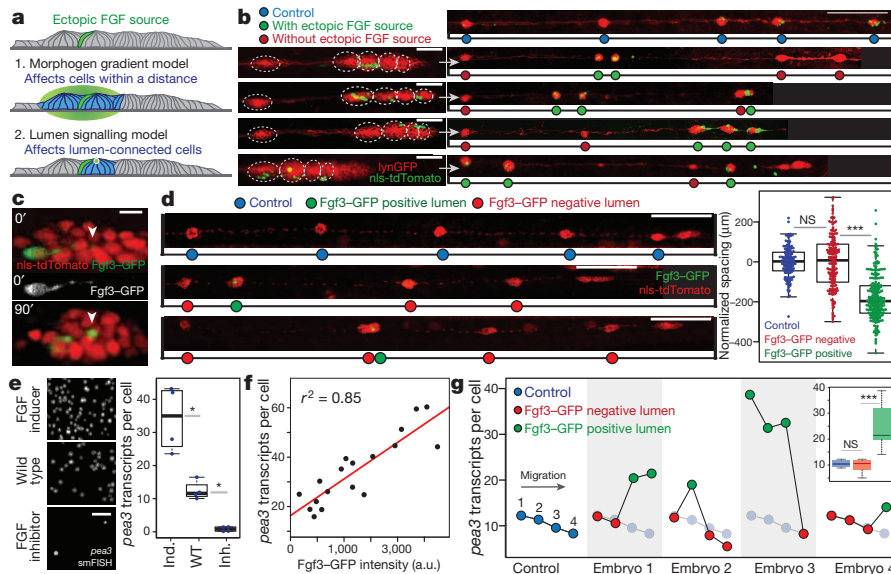
**Figure 2 | Secreted FGF becomes concentrated in multicellular microlumina at the centre of organ progenitors.** Fgf3-GFP and secGFP images are taken from stable LexOP lines unless otherwise stated. **a**, Fgf3-GFP localizes to apical spheres (arrowheads) after uniform expression (red membrane counter-label, cells outlined in dashed lines in side view). **b**, Kymographs showing that the appearance of Fgf3-GFP and secGFP accumulation correlates with onset of organ deposition. **c**, secGFP localization in primordium. **d**, Intracellular accumulation of Fgf3-GFP vesicles after brefeldin A (+BfA) treatment. **e**, Schema comparing expression (green) and localization (yellow) patterns of Fgf3-GFP. **f**, Fgfr1a protein distribution (anti-Fgfr1a antibody) in migrating primordium. **g**, Anti-Fgfr1a antibody staining in apical and basal optical slices of a rear rosette (calibration bar: signal intensity). **h**, **i**, Tight junction 'bucky ball' in deposited organs (yellow) and tight

junctions of overlying skin cells (blue). **i**, Higher-resolution view of **h**. **j**, Co-labelling of Fgf3-GFP (green) and tight junctions (ZO1, red). **k**, BAC *fgf3:Fgf3-GFP* embryo shows microluminal localization, membranes labelled with BAC *cxc4b:Cxcr4b-RFP* (RFP, red fluorescent protein). **l–n**, CLEM analysis of microlumen. **l**, Fluorescent images used for CLEM alignment showing slice position (dashed), orthogonal view (below), secGFP and electron microscope overlay at same position (right). **m**, Plasma membrane tracking shows microlumina is assembled from many cells (pseudocoloured). **n**, Microluminal space and surrounding cell junctions (schematic: section position), boxes show higher-resolution view (red, tight junctions (TJ); orange, adherens junctions (AJ)). **o**, Schematic representation of organ with central microlumen. Scale bars, 50  $\mu$ m (**a**, **c**, **d**), 200  $\mu$ m, 5 h (**b**), 20  $\mu$ m (**f**), 5  $\mu$ m (**g**, **h**, **i**, **j**, **k**, **l**, **m**, **n**).

of Fgf3-GFP revealed that it was concentrated in spherical volumes at the apical centre of organ progenitor rosettes (Fig. 2a, e). Time-lapse analysis showed that the appearance of these spheres correlated with the deceleration and arrest of the associated organ progenitor (Fig. 2b and Supplementary Video 6). Apical spheres were also observed when the tissue expressed a secreted form of GFP (secGFP; Fig. 2b, c). Inhibiting protein secretion with brefeldin A prevented Fgf3-GFP localization to these apical spheres and retained it in vesicles within all cells of the primordium (Fig. 2d), indicating that these spheres represent tightly restricted pools of apically secreted proteins (Extended Data Fig. 4). By contrast, direct visualization of endogenous Fgfr1a, using a newly generated monoclonal antibody against the zebrafish protein, revealed an unrestricted plasma membrane distribution of the receptor (Fig. 2f, g). Immunofluorescence of tight-junctions<sup>17</sup> (Fig. 2h–j) and ultrastructural analysis using correlative light electron microscopy (CLEM; Fig. 2l–n and Extended Data Fig. 5)<sup>18</sup> demonstrated that these apical spheres of secreted protein represent extracellular pockets, or microlumina, assembled from cell apical domains and displaying the cell junctions characteristic of a lumen (Fig. 2n, o and Extended Data Fig. 5). Identical luminal localization was also observed when Fgf3-GFP was expressed at normal physiological levels using BAC-mediated complementation (*fgf3:fgf3-GFP*; Fig. 2k and Extended Data Fig. 6). Interestingly, these microlumina showed a geodesic organization to which each cell of the organ progenitor contributes a facet and thus has access to this shared microenvironment (Fig. 2m, o). The secGFP signal correlated perfectly with the shape of the luminal cavity, even filling 'side-pockets' that are formed stochastically by protruding sensory kinocilia of differentiating

organs (Fig. 2l, Extended Data Fig. 5 and Supplementary Video 7), suggesting that secreted proteins freely diffuse within the microlumen. Indeed, fluorescence loss in photobleaching (FLIP) and fluorescence recovery after photobleach (FRAP) analysis of Fgf3-GFP confirmed that Fgf3-GFP is highly mobile within the microlumen (Extended Data Fig. 5).

The results described above reveal that microlumina could act as 'hubs' that locally concentrate secreted FGF molecules and ensure coordinated signalling responses within the migrating tissue. Alternatively, signalling activity may be determined by concentration gradients of freely diffusible FGF molecules in the open extracellular environment, consistent with its known role as a morphogen in other contexts<sup>19</sup>. To distinguish between these two models (Fig. 3a), we investigated the range of FGF action by overexpressing the protein from randomly positioned cell clones that were generated either by cell transplantation (Fig. 3b and Extended Data Fig. 7) or mosaic expression of *lexOP:fgf3-GFP* (Fig. 3d). Interestingly, Fgf3-GFP was secreted into microlumina independently of the position of the expressing cell within the rosette, indicating that any cell of the group can contribute signal to the microluminal pool (Fig. 3c and Extended Data Fig. 7). As a first readout of FGF activity, we mapped the deposition intervals of organs with and without ectopic Fgf3-GFP-expressing cells. As shown in Fig. 3, individual ectopic Fgf3-GFP-expressing cells efficiently arrested the migration of cells that were connected to the same Fgf3-GFP-positive microlumen but they had no effect on cells in neighbouring organs, even when they were physically closer than cells of the same organ (Fig. 3b, d, Extended Data Fig. 7 and Supplementary Video 8). Thus, the organ deposition response to ectopic



**Figure 3 | Microlumina focus FGF-signalling activity within migrating collective.** **a**, Schema showing possible outcomes of ectopic FGF experiments. **b**, *lexOP:Fgf3-GFP/cxcr4b:nls-tdTomato* cell clones (green) transplanted into *cldnb:lynGFP* primordium (red), showing future organ territories (dashed line) and final pattern of organ deposition (right). **c**, Time-lapse of microluminal filling by single Fgf3-GFP cell after drug induction in mosaic tissue (arrow heads pointing organ centre). **d**, Organ deposition pattern after mosaic *lexOP:fgf3-GFP* expression by transient injection. Box-plot of organ spacing in mosaic embryos with Fgf3-GFP-positive microlumen (green,  $n = 240$ ), without Fgf3-GFP-positive microlumen (red,  $n = 170$ ) and control

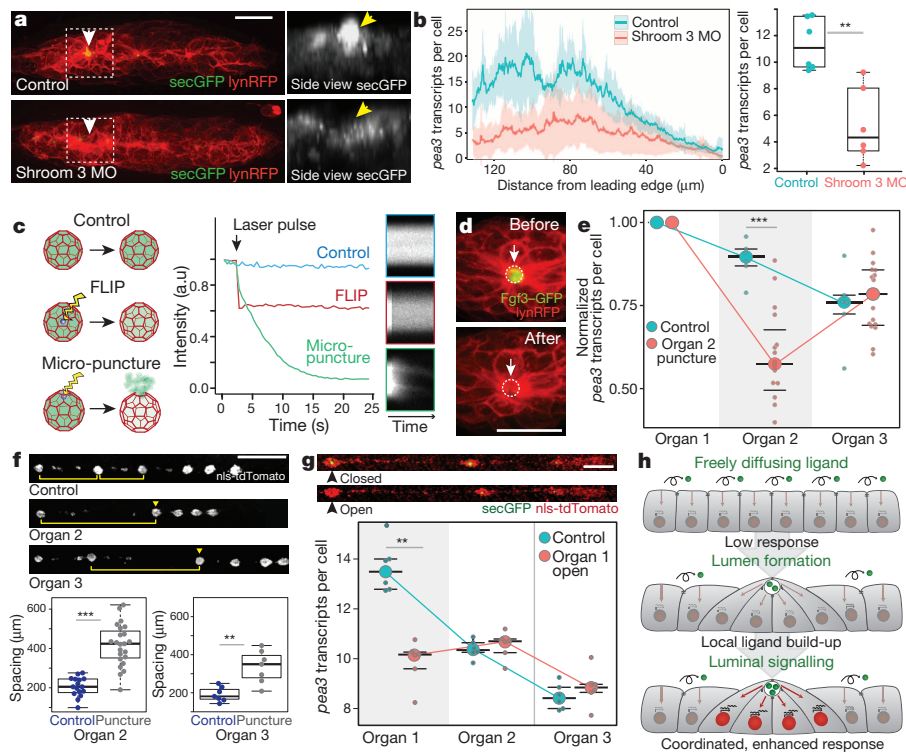
FGF activity was local and coordinated, affecting only cells that shared a microlumen with Fgf3-GFP source cells. To measure FGF signalling more directly, we next monitored the transcription of its immediate target gene *pea3* (ref. 20) by single-molecule fluorescent *in situ* hybridization approach (smFISH), which revealed a clear Fgf-dependent response (Fig. 3e, f, Extended Data Fig. 8 and Supplementary Video 9). *Pea3* smFISH analysis of primordia carrying ectopic Fgf3-GFP-expressing clones showed that all cells in contact with Fgf3-GFP-positive microlumina upregulated target gene transcription, whereas cells from neighbouring rosettes showed no change (Fig. 3g), confirming that the response to FGF signalling is highly restricted. In addition, this revealed that upregulated FGF signalling responses were uniform across individual organ precursors. In conclusion, single-cell ectopic expression of FGF does not support a model where FGF responses are determined by a concentration gradient diffusing from secreting cells. By contrast, these data are fully consistent with the model where the FGF signal is locally concentrated, and collectively presented, by a central microlumen. Thus, the formation of microlumen allows FGF signalling, widely studied for its role in mediating long-range positional information<sup>19</sup>, to coordinate locally the behaviour of discrete cell groups within the migrating tissue.

The luminal signalling model predicts that these multicellular assemblies are required for efficient FGF signalling responses. To test this, we first applied a genetic method to prevent microluminal assembly, an approach complicated by the fact that many key regulators of this process also control epithelial polarity<sup>21,22</sup>. We therefore knocked down shroom3, an actin-binding<sup>23</sup> protein required for apical constriction of organ progenitor cells but not for epithelial polarity<sup>24</sup>. Morpholino knock-down of shroom3 transiently suppressed microluminal formation, as revealed by a failure to concentrate secGFP (Fig. 4a). smFISH analysis of *pea3* confirmed that FGF target gene transcription was significantly reduced (Fig. 4b). Second, we acutely opened microlumen structures by two-photon laser micropuncture, which caused rapid leakage of Fgf3-GFP (Fig. 4c, d, Extended Data Fig. 9 and Supplementary Video 10), confirming that these local build-ups of FGF signal are dependent on microlumen integrity. The microluminal opening by laser micropuncture

was transient, as revealed by the recovery of characteristic microluminal Fgf3-GFP spheres, showing that this targeted perturbation had negligible effects on cell viability (Extended Data Fig. 9). Nevertheless, smFISH analysis of *pea3* revealed that target gene expression was reduced after transient depletion of microluminal Fgf3-GFP, demonstrating that trapping of secreted FGF is required to maintain high signalling levels (Fig. 4e). In addition, microluminal opening specifically delayed the deposition of targeted organs by prolonging their migration, a direct confirmation that microlumina are required for FGF to exert its biological role during this organogenesis process (Fig. 4f). Finally, we exploited the fact that the microluminal cavity is opened when deposited organ progenitors fuse with the overlying skin, a natural event that also leads to rapid loss of Fgf3-GFP (Extended Data Fig. 10). Since the timing of skin fusion varies between embryos, we could directly compare organs where microlumina had just opened with those that were still closed at identical developmental stages (Fig. 4g, Supplementary Video 6). smFISH analysis of *pea3* in this unperturbed context revealed that FGF signalling was again reduced specifically in organs with opened microlumina (Fig. 4g). Combined, these data provide compelling experimental support for a model where microlumina act as shared microenvironments that locally concentrate FGF to enhance signalling within the migrating tissue (Fig. 4h).

Previous studies addressing the regulation of extracellular signals have focused on the role of additional cell surface or extracellular proteins, such as heparan sulphate proteoglycans<sup>25</sup> and receptors<sup>26</sup>, whose own spatiotemporal regulation is currently under investigation<sup>27</sup>. Here, we uncover an alternative mechanism that instead exploits an intrinsic biological feature of epithelial tissues, namely their ability to assemble a shared enclosed lumen<sup>6</sup>. This finding has important implications for understanding how responses to extracellular signals are controlled and coordinated in tissues *in vivo*. To our knowledge, it provides the first such mechanism that acts specifically at the level of multicellular organization, as only cell groups that assemble a central lumen are able to trap and concentrate the freely diffusible ligand (Fig. 4h). We propose that formation of the microlumen is required to restrict, coordinate and





**Figure 4 | Microluminal assembly and integrity are required for efficient FGF signalling.** **a, b,** Knockdown of shroom3 (Shroom3 MO) prevents microlumen formation and leads to lack of apical spheres in *lexOP:secGFP* primordia (**a**, arrowhead), and reduces *pea3* transcription, shown in profile plot and box-plot quantification (**b**,  $N = 6, 6$ ). **c–e,** Schema of two-photon micropuncture approach. **c,** Schematic representation of micropuncture experiment (left). Plot of Fgf3-GFP pool fluorescence intensity after micropuncture (green) or internal bleach pulse (red). Panels show kymographs. **d,** Images of Fgf3-GFP-positive organ before and after micropuncture. **e,** *Pea3* smFISH of Fgf3-GFP-expressing organs 30–60 min after micropuncture, comparing organ 2 micropunctured samples (pink,  $N = 16$ ) with controls (blue,  $N = 6$ ). Absolute *pea3* transcripts per cell were normalized to first organ for each embryo. **f,** Organ deposition delay after luminal micropuncture of Fgf3-GFP-expressing second ( $n = 16, 23$ ) and third ( $n = 7, 7$ ) organs. Quantification of organ deposition spacing (bottom). **g,** Comparison of secGFP- (green) expressing primordia (red) in identical stage embryos; lower specimen shows loss of secGFP microlumina in organ 1 after fusion with the overlying skin (arrowhead). smFISH reveals reduced *pea3* transcript levels in opened organs when compared with unopened organs ( $N = 6, 6$ ). **h,** Schematic representation of microluminal signalling model. Scale bars, 20  $\mu\text{m}$  (**a, d**), 200  $\mu\text{m}$  (**f**), 100  $\mu\text{m}$  (**g**). Statistics: Wilcoxon. \*\* $P < 0.01$ ; \*\*\* $P < 0.001$ .

enhance FGF signalling within the migrating tissue. This local increase in FGF activity then positively feedbacks on microlumina by increasing the epithelial character of responding cells, leading to the formation of stable rosettes. Thus, luminal hubs provide a morphogenetic checkpoint function by ensuring, in this case, that cells become polarized and organized before they can respond efficiently to signals promoting their differentiation. Moreover, as lumen formation itself is highly sensitive to changes in epithelial polarity and adhesion<sup>21,28</sup>, it is likely that luminal signalling hubs can be rapidly disassembled and reassembled by processes that alter cell cohesion, such as the epithelial–mesenchymal transition that is a hallmark of organogenesis and cancer<sup>29</sup>. However, this mechanism could potentially be active in any context where cells construct a lumen or similar enclosed extracellular microenvironment, such as the transient tissue-folds that are prevalent during morphogenesis<sup>30</sup>. A notable example is provided by the recent finding that early mammalian embryos and embryonic stem cells self-organize to form polarized rosettes with a central lumen<sup>7</sup>, structures that are morphologically highly similar to those interrogated here. Our study suggests potential signalling roles for shared lumina in many other tissue contexts.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 15 April; accepted 9 September 2014.

Published online 22 October 2014.

- Sasai, Y. Cytosystems dynamics in self-organization of tissue architecture. *Nature* **493**, 318–326 (2013).
- Sato, T. & Clevers, H. Growing self-organizing mini-guts from a single intestinal stem cell: mechanism and applications. *Science* **340**, 1190–1194 (2013).
- Eiraku, M. *et al.* Self-organizing optic-cup morphogenesis in three-dimensional culture. *Nature* **472**, 51–56 (2011).
- Lancaster, M. A. *et al.* Cerebral organoids model human brain development and microcephaly. *Nature* **501**, 373–379 (2013).
- Ghyssen, A. & Dambly-Chaudière, C. The lateral line microcosmos. *Genes Dev.* **21**, 2118–2130 (2007).
- O'Brien, L. E., Zegers, M. M. P. & Mostov, K. E. Opinion: building epithelial architecture: insights from three-dimensional culture models. *Nature Rev. Mol. Cell Biol.* **3**, 531–537 (2002).
- Bedzhov, I. & Zernicka-Goetz, M. Self-organizing properties of mouse pluripotent cells initiate morphogenesis upon implantation. *Cell* **156**, 1032–1044 (2014).
- Reeves, G. T., Muratov, C. B., Schüpbach, T. & Shvartsman, S. Y. Quantitative models of developmental pattern formation. *Dev. Cell* **11**, 289–300 (2006).
- Metzger, R. J., Klein, O. D., Martin, G. R. & Krasnow, M. A. The branching programme of mouse lung development. *Nature* **453**, 745–750 (2008).
- Revenu, C. *et al.* Quantitative cell polarity imaging defines leader-to-follower transitions during collective migration and the key role of microtubule-dependent adherens junction formation. *Development* **141**, 1282–1291 (2014).
- Ma, E. Y. & Raible, D. W. Signaling pathways regulating zebrafish lateral line development. *Curr. Biol.* **19**, R381–R386 (2009).
- Lecaudey, V., Cakan-Akdogan, G., Norton, W. H. J. & Gilmour, D. Dynamic Fgf signaling couples morphogenesis and migration in the zebrafish lateral line primordium. *Development* **135**, 2695–2705 (2008).
- Nechiporuk, A. & Raible, D. W. FGF-dependent mechanosensory organ patterning in zebrafish. *Science* **320**, 1774–1777 (2008).
- Aman, A. & Piotrowski, T. Wnt/ $\beta$ -catenin and Fgf signaling control collective cell migration by restricting chemokine receptor expression. *Dev. Cell* **15**, 749–761 (2008).
- Rohner, N. *et al.* Duplication of *fgfr1* permits Fgf signaling to serve as a target for selection during domestication. *Curr. Biol.* **19**, 1642–1647 (2009).
- Emelyanov, A. & Parinov, S. Mifepristone-inducible LexPR system to drive and control gene expression in transgenic zebrafish. *Dev. Biol.* **320**, 113–121 (2008).
- Bagnat, M., Cheung, I. D., Mostov, K. E. & Stainier, D. Y. R. Genetic control of single lumen formation in the zebrafish gut. *Nature* **9**, 954–960 (2007).
- Kolotuev, I., Schwab, Y. & Labouesse, M. A precise and rapid mapping protocol for correlative light and electron microscopy of small invertebrate organisms. *Biol. Cell* **4**, 121–132 (2009).
- Yu, S. R. *et al.* Fgf8 morphogen gradient forms by a source-sink mechanism with freely diffusing molecules. *Nature* **461**, 533–536 (2009).
- Raible, F. & Brand, M. Tight transcriptional control of the ETS domain factors *Ern* and *Pea3* by Fgf signaling during early zebrafish development. *Mech. Dev.* **107**, 105–117 (2001).
- Martin-Belmonte, F. *et al.* Cell-polarity dynamics controls the mechanism of lumen formation in epithelial morphogenesis. *Curr. Biol.* **18**, 507–513 (2008).
- Kesavan, G. *et al.* Cdc42-mediated tubulogenesis controls cell specification. *Cell* **139**, 791–801 (2009).
- Hildebrand, J. D. Shroom regulates epithelial cell shape via the apical positioning of an actomyosin network. *J. Cell Sci.* **118**, 5191–5203 (2005).
- Ernst, S. *et al.* Shroom3 is required downstream of FGF signalling to mediate proneuromast assembly in zebrafish. *Development* **139**, 4571–4581 (2012).

25. Belenkaya, T. Y. *et al.* *Drosophila* Dpp morphogen movement is independent of dynamin-mediated endocytosis but regulated by the glypican members of heparan sulfate proteoglycans. *Cell* **119**, 231–244 (2004).
26. Donà, E. *et al.* Directional tissue migration through a self-generated chemokine gradient. *Nature* **503**, 285–289 (2013).
27. Bökel, C. & Brand, M. Endocytosis and signaling during development. *Cold Spring Harb. Perspect. Biol.* **6**, a016881 (2014).
28. Bryant, D. M. & Mostov, K. E. From cells to organs: building polarized tissue. *Nature Rev. Mol. Cell Biol.* **9**, 887–901 (2008).
29. Nieto, M. A. Epithelial plasticity: a common theme in embryonic and cancer cells. *Science* **342**, 1234850 (2013).
30. Wang, Y.-C., Khan, Z., Kaschube, M. & Wieschaus, E. F. Differential positioning of adherens junctions is associated with initiation of epithelial folding. *Nature* **484**, 390–393 (2012).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** We are grateful to J. Ellenberg, S. de Renzis and F. Peri for suggestions and comments on the manuscript, A. Aulehla for advice about timing, and E. Karsenti and the Gilmour laboratory for discussion. We thank M. Brand for advice

about FGF tagging, the EMBL Advanced Light Microscopy Facility, in particular Y. Belyaev, for imaging assistance, the European Molecular Biology Laboratory (EMBL) Monoclonal Antibody (MACF) and Protein Expression Facilities for Fgfr1a antibody, K. Miura from the EMBL Centre for Cell and Molecular Imaging for advice with data analysis, E. Dona and T. Gregor for advice with the smFISH protocol, and A. Gruia for fish care. We acknowledge funding from the European Molecular Biology Organization and EMBL Interdisciplinary Postdocs (EIPD) (to C.R.) and the Deutsche Forschungsgemeinschaft SFB 488 (to D.G.).

**Author Contributions** D.G. and S.D. designed the study. S.D. performed all experiments, with the exception of CLEM experiments performed with N.S. and Y.S., and antibody-based analysis of the microlumen performed by C.R. S.D. and M.I. developed the data analysis methods with input from P.B. A.K. developed the LexPR inducible gene expression system and C.R. generated the Cxcr4b-RFP line. D.G. and S.D. interpreted the data and wrote the paper with input from all authors.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.G. ([gilmour@embl.de](mailto:gilmour@embl.de)).

## METHODS

**Fish handling.** Zebrafish (*Danio rerio*) strains were maintained following standard protocols<sup>31</sup>. Embryos were raised in E3 buffer at 26–30 °C. All zebrafish experiments were conducted on embryos younger than 3 d.p.f., under the rules of the European Molecular Biology Laboratory and the guidelines of the European Commission, Directive 2010/63/EU. In all experiments involving chemical treatment, embryos were dechorionated with pronase. Before live imaging and fixation procedures, embryos were anaesthetized with 0.01% tricaine. For *in situ* hybridization and immunostaining experiments, embryos were treated with 0.002% phenylthiourea at 24 hours post-fertilization (h.p.f.) to prevent pigmentation. For live imaging, embryos were mounted in 0.8% low melting agarose in glass-bottom dishes (MatTek or CELLview) and imaged at 28 °C unless otherwise stated. The following mutant and transgenic strains were used: *fgfr1a*<sup>13R705H</sup> (ref. 15), *cxcr4b:nl5-tdTomato*<sup>26</sup>, *cldnb:lynGFP*<sup>32</sup>.

**Inducible gene expression system and BAC lines.** The LexPR/LexOP transactivation system<sup>16</sup> was used to express genes in the lateral line upon addition of the progesterone analogue RU486. Two LexPR 'driver' lines were generated by inserting the cassettes (1) LexPR/polyA/LexOP:lynRFP/SV40polyA/FRT-KanR-FRT or (2) LexPR/SV40polyA(L)/FRT-KanR-FRT into the first exon of the *Cxcr4b* BAC clone CH211-145M5 by ET recombineering (Gene Bridges). The KanR cassette was subsequently removed with FLP recombinase. To serve as a transgenic marker, the 'crystal eye' cry:CFP/KanR cassette was inserted in the BAC backbone as a transgenic marker. Modified BACs were purified (Large Construct Kit, Qiagen) and injected into one-cell stage embryos to generate transgenic driver lines. The following LexOP 'responder' lines were generated from multisite-gateway clones (Invitrogen) using the Tol2kit<sup>33</sup>: (1) LexOP(p5E)/secGFP(pME)/fgf3(p3E), (2) LexOP(p5E)/secGFP(pME)/polyA(p3E) and (3) LexOP(p5E)/nlsGFP(pME)/polyA(p3E). All clones carry the *cmlc2:eGFP* 'bleeding heart' cassette as a transgenic marker<sup>33</sup>. Fgf3 was tagged by inserting the GFP sequence in between signal peptide (sec) and globular domain of Fgf3, the strategy previously used for Fgf8 (ref. 19). secGFP has the signal peptide of Fgf3 protein fused to GFP. The LexPR system was proved to be non-leaky as there was no detectable target gene expression in the absence of activator RU486. It was also proved to be spatially restricted, showing transactivation only in *Cxcr4b*-expressing tissues. The dose–response of the inducible LexPR system was evaluated using the transactivation levels of the *lexOP:nlsGFP* responder after treatment for 6 h with 0, 5, 10 and 20 µM RU486. The lateral line primordium was imaged with the same acquisition settings for all samples. The average fluorescence signal from mean projected images was quantified, and these values were first background subtracted (calculated from untreated embryos) and then normalized to [0,1] range by dividing with the maximum signal.

The BAC *fgf3:fgf3-GFP* line was generated by replacing coding sequence of the first exon in Fgf3 BAC (CH211-96B20) with a targeting cassette: secGFP/fgf3/SV40polyA(FRT-KanR-FRT). The KanR cassette was subsequently removed with FLP recombinase. The cry:CFP/Ampr 'crystal eye' cassette was inserted into the BAC backbone as a transgenic marker<sup>26</sup>. BAC recombination and purification steps were followed as described<sup>26</sup>. Whole-embryo overview images were generated using a Zeiss Lightsheet Z.1 microscope (×20). The functionality of the BAC line was tested by FGF knockdown rescue experiments, where both Fgf3 and Fgf10a genes were knocked down owing to mutual compensation of the two ligands in lateral line system<sup>12</sup>. Fgf3 (splice site blocker, 5 ng nl<sup>-1</sup>)<sup>34</sup> and Fgf10 (start site blocker, 5 ng nl<sup>-1</sup>)<sup>35</sup> morpholinos were injected into BAC *fgf3:fgf3-GFP*, *cldnb:lynGFP* embryos at the one-cell stage, where half of the embryos were BAC *fgf3:fgf3-GFP* transgene carriers as detected by cry:eCFP transgenic marker (Extended Data Fig. 6). The *cxcr4b:cxcr4b-tagRFP* line was generated by inserting TagRFP cassette into the *Cxcr4b* BAC (CH211-145M5) as described<sup>26</sup>.

**Chemical treatments.** SU5402 (Calbiochem) was used for inhibition of Fgfr1 kinase activity. Eight embryos per 2 ml of E3 buffer were used as standard treatment density. For organ deposition experiments, embryos were treated with 0.5 and 1 µM SU5402 in 0.1% dimethylsulphoxide (DMSO) starting at 24 h.p.f.; controls were treated with 0.1% DMSO alone. Time-lapse imaging was started at 4–6 h after treatment, when the first organ was about to be deposited. Drug efficacy was observed to decrease over time-lapse imaging owing to light sensitivity of SU5402. Therefore, organ spacing was quantified as 2 d.p.f. measurements on embryos that were kept in the dark.

RU486 (Sigma) was used to transactivate LexPR/LexOP driven gene expression. For organ deposition experiments, *cxcr4b:lexPR*, *lexOP:fgf3-GFP* embryos were treated with 5, 10 and 20 µM RU486 starting at 24 h.p.f. Time-lapse imaging was started at 4–6 h after treatment. As a control group, *cxcr4b:LexPR* transgenics without *lexOP:fgf3-GFP* were treated with 10 µM RU486.

Brefeldin A (BFA, Sigma) was used to visualize localization of Fgf3-GFP in the absence of secretion. Embryos were first treated with 15 µM RU486 for 4–6 h to express Fgf3-GFP and then treated with 14 µM BFA for 30 min to block secretion.

**Immunofluorescence and colorimetric *in situ* hybridization.** Monoclonal anti-Fgfr1a antibody was generated using the following peptide, corresponding to 140–360 amino acids of Fgfr1a protein as an antigen: 'KLSNDQNLPMAPVWAQPKMEKKLHAPASKTVKFRQANGNPTTLKWLKNGKEFKRDQRIGGFKVREHMWTIMEVPSDRGNYTCLVENRHGSINHTYQLDVVERSHPRPILQAGLPANRTAVVGSDFEVECKVFSDPQPHIQWLKHIEVNGSRYGPDGLPYV RALKTAGVNTTDKEMEVLQIRNVSLDAGEYTCLAGNSIGHSHHSAWLTV YKA'. For whole mount antibody staining, embryos were fixed with pre-cooled 85% methanol, 15% acetic acid for 3 min at –20 °C and rehydrated with methanol series 75, 50, 25%, 3 min each at room temperature (~23 °C). Blocking was done with blocking buffer (1× PBS, 1% DMSO, 2% NCS, 1% BSA, 0.1% Tween) for 4 h at room temperature. Embryos were then incubated with primary antibody (1:50 in blocking buffer) for 20 h at 4 °C. Embryos were washed with blocking buffer four times for 30 min at room temperature and incubated with Alexa 488-anti-mouse antibody (1:500 in blocking buffer) for 2.5 h at room temperature. Embryos were then washed with blocking buffer four times for 30 min and mounted in 1% LM agarose. Samples were imaged using an Ultraview VoX spinning disk confocal microscope with a ×63 Zeiss water objective (1.2 numerical aperture).

ZO1 antibody staining (anti-ZO1 primary antibody, Alexa-568-coupled anti-mouse secondary antibody) and *pea3* and *fgf3* *in situ* hybridization (DIG probes, anti-DIG alkaline phosphatase coupled antibody, NBT/BCIP substrate at 30 °C) were performed as described previously<sup>12</sup>.

**Analysis of migration and organ patterning.** Embryos were imaged with PE Ultraview ERS and PE Ultraview VoX spinning disk microscopes using Zeiss ×5, ×10, ×20 air objectives. Multi-position time-lapse images were acquired from 10 to 30 min intervals and a computational pipeline to analyse migration and organ patterning was established. First, images of individual embryos were stitched automatically by a macro using Grid stitching tool in FIJI<sup>36</sup>. To analyse time-lapse movies, kymographs (*x*–*t* graphs) were generated using FIJI. For each embryo, a segmented line region of interest (ROI) was drawn along the migration path of the primordium with a thickness that covered the lateral line primordium. The image beneath the line ROI was re-sliced (from *xy*–*t* to *xt*–*y*) and maximum projected. This way maximum signal intensity along the width of the tissue was represented in the kymograph for each time point. Images were then saved as text images to be automatically processed with an R script. Organ positions were determined with a peak detection algorithm implemented in R package 'Peaks'<sup>37</sup>. To map the trajectory of each organ, kymographs were sequentially processed in reverse order from the last time frame to the first. A wide range of parameters (threshold from 10 to 50 in increments of 10, and sigma from 3 to 9 in increments of 3) was used for peak detection, since the signal intensity profiles change over time. We manually checked whether the peaks identified at the last frame referred to an organ or not; only those that did refer were retained as starting points for tracking. For the remaining time points, the hypothetical position of each organ was initially estimated on the basis of the average displacement of the last three time points, then the closest position was sequentially linked to the trajectory between consecutive time points (Supplementary Video 2). Velocity and acceleration profiles of organs were generated from the migration trajectories using local polynomial fitting and its derivatives (KernSmooth package in R)<sup>38</sup>. Organ deposition was defined as the time point where acceleration of the individual organ unit was minimum. We defined three potential parameters that influence organ patterning: (1) embryo growth, (2) primordium migration velocity and (3) organ deposition timing. As higher growth rate between two organ depositions could hypothetically result in increased spacing, we evaluated the effect of embryonic growth by generating trajectories of manually segmented myotome borders, as embryonic landmarks, from kymographs generated using transmission light images. Next, myotome trajectories were subtracted from lateral line organ trajectories using the closest myotome for each organ and each time point. Finally, 'growth-subtracted' organ positions were calculated, revealing that in the absence of embryonic growth, organ spacing would decrease overall without much effect on relative spacing. We next evaluated the effect of primordium migration velocity and organ deposition timing on organ spacing. If primordium velocity was higher between two organ depositions, or the following organ was deposited later, the spacing between these organs would increase. Correlation of these two parameters with spacing revealed that organ deposition timing is the main determinant of the global organ patterning in WT embryos.

**CLEM.** SecGFP, nls-tdTomato-expressing embryos were live imaged (sagittal plane) with a confocal microscope using a ×10 objective for whole-embryo overviews, to aid tissue sectioning, and ×63 objective for high-resolution imaging of lateral line organs, to aid three-dimensional CLEM image construction. After live imaging, embryos were removed from agarose, anaesthetized and tails were removed by cutting after the yolk extension. Bodies were immediately fixed with 2.5% glutaraldehyde and 4% paraformaldehyde in 0.1 M PHEM buffer for 14 min in a Pelco BioWave microwave containing ColdSpot (100 W cycling intervals of 2 min on and off under vacuum). Further processing was performed as described<sup>18</sup>, although using 0.1 M



PHEM buffer instead of cacodylate buffer. Samples were flat embedded between aclar sheets and polymerized at 60 °C for 48 h. Lateral line organs were targeted for further processing by overlaying whole-embryo overviews of live imaging and images of fixed-embedded samples (CLEM targeting approach as described elsewhere<sup>18</sup>). Melanocytes were used as landmarks to correlate the two data sets and then as guides to laser etch the block surface with an Olympus Cell<sup>^</sup>R with UV Cutting. Serial sections were cut 70 nm thick along the dorsoventral axis of the embryo (transverse plane) and placed on a copper palladium slot grid, coated with 1% Formvar (Serva).

Electron microscope imaging was performed on a CM120 Phillips electron microscope. Serial images were aligned with Adobe Photoshop and structures of interest were tracked manually in 3dmod<sup>39</sup>. Electron and fluorescence microscopy images were further processed in Imaris 7.6.4 (Bitplane) for three-dimensional image handling. Nucleus positions of the target organs in electron microscopy images (dark grey) and fluorescence images (nls-tdTomato) were compared using the oblique slicer tool in Imaris to identify the correct transversal sectioning angle. Fluorescence images were then re-sliced using the identified angle. Shrinkage of electron microscope samples was calculated by comparing three-dimensional tissue size in electron microscope images and fluorescence images, then fluorescence images were resized accordingly. The central slice of electron microscope images and the corresponding fluorescence image were overlaid as shown in Extended Data Fig. 4m. A three-dimensional CLEM image construction of an organ centre with segmented structures is displayed in Supplementary Video 7.

**FRAP and FLIP.** Photo-bleaching experiments were performed using an Ultraview VoX spinning disk microscope equipped with a photokinesis unit and Zeiss ×63 water objective. Position accuracy of the laser pulse was calibrated using green fluorescent slides before each experiment. Experiments were performed on middle confocal planes of secGFP and Fgf3–GFP pools. In FLIP experiments, five pre-bleach images were acquired (0.018 s per frame), then a small region (spot ROI with 0.73 µm diameter) was repetitively bleached (45 time points) and the sample imaged in between (0.3 s per frame). Images were analysed by measuring mean intensity over time of (1) bleached region, (2) total pool, (3) background and (4) multiple other regions within the pool. In FRAP experiments, five pre-bleach images were acquired (30 ms per frame), then a strip ROI on the edge of the pool was bleached once and post-bleach images were acquired (45 time points, 30 ms per frame). Images were analysed by measuring mean intensity over time of (1) bleached region, (2) total pool and (3) background. Next, the measurements were uploaded to easyFRAP to calculate half-time of recovery with full-scale normalization and double term fitting<sup>40</sup>. Small spot ROI bleaching in the centre of the pool could not be used for FRAP experiments as the redistribution of the protein was too fast to catch recovery curves. This fast distribution could also be seen by the FLIP experiments with a spot ROI bleaching.

**Secretory pathway analysis.** *In vitro* synthesized messenger RNAs (mRNAs) (100 ng µl<sup>-1</sup>) encoding GM130-tdTomato and KDEL peptide fused to mKate2 were injected into one-cell stage embryos to label the Golgi apparatus and endoplasmic reticulum, respectively. GM130-tdTomato signal was segmented in three-dimensions and used as a landmark for density profile plotting of secGFP and Fgf3–GFP intensities within each cell.

**Single-cell overexpression experiments.** Fgf3–GFP mis-expressing cell clones were generated by cell transplantation, following established protocols. Donor cells from *cxc4b:lexPR*, *lexOP:lynRFP*, *lexOP:Fgf3–GFP*, *cxc4b:nls-tdTomato* transgenic embryos were transplanted into *clnbn:lynGFP* transgenic embryos, allowing Fgf3-expressing clones to be marked with nuclear tdTomato in membrane GFP-labelled hosts. *Cxc4b:nls-tdTomato* cells were transplanted into *clnbn:lynGFP* embryos as controls. Time-lapse imaging was performed and the effect of FGF mis-expression was analysed by comparing migration behaviour of organs with and without clones. Primordium velocity, organ spacing and organ deposition timing between two consecutive depositions were quantified. To correct for intrinsic variation, which is high among WT organ intervals (see Fig. 1a and Extended Data Fig. 1), calculated values for each interval were normalized to the mean of the corresponding interval from controls.

To generate Fgf3–GFP-overexpressing clones by mosaic expression, the *lexOP:Fgf3–GFP* plasmid was injected into *cxc4b:lexPR*, *cxc4b:nls-tdTomato* transgenic embryos at the one-cell stage. The next day, Fgf3–GFP expression was observed in randomly positioned cells. Overview images of the lateral line were acquired at 2 d.p.f. to analyse organ patterning; water injected embryos were used as controls. Organ spacing in Fgf3–GFP microlumina-positive and -negative organs was analysed by normalizing each interval to control embryos as described for the transplantation experiment above.

**smFISH.** smFISH probes (Custom Stellaris FISH probes, Biosearch Technologies) were designed to target *pea3* mRNA (ENSDART0000013033). Forty-eight sequence-specific oligonucleotides (listed below) were conjugated to the fluorophores Cal Fluor 590 (red) and Quasar 670 (far red). Embryos were fixed and permeabilized following standard zebrafish *in situ* hybridization protocols. smFISH was performed

following the protocol of ref. 41, with the exception that 5× SSC replaced 2× SSC in the hybridization buffer, and embryos were stained with DAPI for 15 min at 30 °C after probe removal. Embryos hybridized with Cal Fluor 590 conjugated probes were mounted in Aquamount (Polysciences). Embryos hybridized with Quasar 670 conjugated probes were mounted in GLOX buffer (0.4% glucose, 10 mM TrisHCl (pH 8), 2× SSC, 0.16 mg ml<sup>-1</sup> glucose oxidase, 0.02 mg ml<sup>-1</sup> catalase in ddH<sub>2</sub>O) and imaged immediately to prevent bleaching. Imaging was performed using a ×100 Zeiss oil objective (1.4 numerical aperture) and a PE Ultraview VoX spinning disk microscope with 0.07 µm pixel size and 0.2 µm z steps. For Cal Fluor 590 conjugated probes, 561 nm excitation, 620(W60) emission, and for Quasar 670 conjugated probes, 640 nm excitation, 705(W90) emission, were used.

smFISH images were analysed in Imaris 7.6.4 (Bitplane). First, a volume of interest (surface object) was defined by manually tracking borders at multiple z slices considering membrane and nucleus labelling ('contour surface' tool). Nuclei were counted using the spot segmentation tool with 2.5 µm estimated diameter, then identified nucleus points were manually corrected for missing or fused selections. The RNA signal was counted using the spot segmentation tool with region growing, local contrast algorithms and 0.4 µm estimated diameter. Identified spots were filtered to have at least 0.4 µm diameter in the z dimension (Supplementary Video 9). Positions of nuclei and RNA spots were exported from Imaris to be processed further in R. Transcript count per cell was calculated by simply dividing the number of identified transcripts by the number of nuclei.

Transcript profiles along the posterior–anterior axis of the primordium were generated by fitting a line (the first principal component) to the nucleus positions along the long axis of the primordium. Then, segmented transcript and nuclei positions were projected on this line and their ratio along the primordium was plotted with 10 µm sliding window. Transcript distributions of a mosaic embryo organ were represented by assigning the transcripts to the closest nucleus position in two dimensions.

To test the validity of the *pea3* smFISH protocol to be used as FGF signalling read-out, embryos were treated with 4 µM SU5402 FGF inhibitor and 15 µM FGF inducer for 6 h. *Pea3* smFISH protocol was applied on four WT, four FGF-induced and four FgfR-inhibited primordia (an average of 150, 143 and 195 cells per primordium respectively), and transcript counts per cell were plotted.

The relation between luminal FGF levels and *pea3* transcription response was tested by inducing *lexOP:Fgf3–GFP* expression with 5 µM and 20 µM inducer for 6 h to generate a wide range of expression levels. Organ 1 of each embryo was imaged using an Ultraview VoX spinning disk confocal microscope and a ×63 Zeiss water objective (1.2 numerical aperture) with the same imaging settings. Embryos were then fixed and processed individually for smFISH protocol to compare their luminal Fgf3–GFP intensity with *pea3* transcript counts.

*Pea3* smFISH oligonucleotides: 1, AAGGAAGACGGACAGAGGCA; 2, CTGTGTTTAAATGAGCTCCA; 3, CTTAACCCTTTGTGGTCATT; 4, CCATCCATCTTAATCCAT; 5, AGTATAAGGCACCTTGCTGGT; 6, ATTCCTTGCGCATATTAG; 7, TCAACAGTCTATTAGGGGC; 8, ATGTATTTCCTTTTGTGCG; 9, AAGAGGTCTTCAGATTCTGT; 10, CCTGAAGTTGGCTTAAATCC; 11, GGAACCTGAGCTTCGGTGA; 12, AACAACTGCTCATCGCTGT; 13, CACTGAGTTCTTGAGTGAA; 14, TTCTTAATCTTCACAGGCGG; 15, TAGCTGAAGCTTCTGTTGTG; 16, TCATAGGCACTGGCGTAAAG; 17, CTGGAATGAGCTCTTAGAT; 18, TTGGGGGAATAATGCTGCAT; 19, TGAGGGTGGATTTCATATACC; 20, CGGAAGGGAACCTGGAAGT; 21, AGAGTGTTCGGATGGAAAC; 22, TGCTGAGGAGGATAAGGCA; 23, CCATGTACTCC TGCTTAAAG; 24, TCCTGTTTGACCATCATATG; 25, CAGGTTCGTAAGTG TAGTCC; 26, TGTGATGGTACATGGATGGG; 27, AAACATGTAGCCTTCACTGT; 28, TGGCACAACACGGGAATCAT; 29, TCACCTCACCTTCAAATTTC; 30, ACCTTCACGAAACACACTGC; 31, TAGTTGAAGTGAGCCACGAC; 32, GAAGGGCAACCAAGAAGTGC; 33, ATGCGATGAAGTGGGCGATTG; 34, ATGAGTTGAATTCATGCG; 35, TTGTCATAGTTTCATGGCTGG; 36, GTAACGAAAGAGGCACTCA; 37, TTTTGCAATAATCCCTTCTC; 38, AGGTTATCAAAGCTTCTGGC; 39, CGCTGATTGTGCGGAAAGC; 40, GTTGACGTAGCGCTCAAATT; 41, AAGAACTCCCTCATCGAGG; 42, TACATGTAGCTTTGGAGTA; 43, AAAGGAGAATGTGGTGGCA; 44, GTGGTAACTGGGATGGGA; 45, ATACAAGAGGATGGGGTGGG; 46, GAATGCAGAGTCCCTAATG; 47, AGATAGGCCTCAGAAGTGAG; 48, GCAATCTCTTGAACCACAGT.

**Shroom3 knockdown.** Shroom3a was knocked down using a previously published morpholino (5'-CCTAATAAATTGTTACCTGACTAAC-3', Gene Tools, 4.2 pmol per embryo)<sup>24</sup>. Consistent with the published report, the effect of knockdown on apical constriction was observed to be transient. To measure *Pea3* levels in the absence of microluminal trapping, only primordia without visible apical constriction were processed further for smFISH analysis.

**Laser micropuncture.** Micropuncture experiments were performed with a Zeiss LSM 780 NLO 2-Photon microscope with a Zeiss ×63 water objective (1.2 numerical aperture). SecGFP and Fgf3–GFP pools were focused, and a laser (two-photon



960 nm laser) pulse was applied on different regions of the pool. A minimal pulse size (ROI diameter) of 0.43  $\mu\text{m}$  was selected to allow lumen opening with undetectable damage to participating cells. Targeting single pulses at these settings to the middle section of the pool resulted in a one-time reduction in total fluorescence signal, which we term FLIP. Targeting identical pulses to the microluminal lattice caused micropuncture, as revealed by a characteristic decay caused by leaking of the GFP pool through time. For mean fluorescence intensity plots, five time points before the pulse and 20 or more time points after the pulse were acquired, at a rate of 1 s per frame. To perform analysis of target gene response (*pea3*) to micropuncture, embryos were treated sequentially over a 30 min session. After the laser surgery, embryos were incubated for 30 min, allowing each organ 30–60 min response time after micropuncture before fixation for smFISH analysis. As there was high intrinsic variability of transactivation using the LexPR system (Extended Data Fig. 2), this resulted in variability in target gene response (Fig. 3). Therefore, to allow direct comparison of transcript counts, values were normalized to the first organ of each embryo that was left unperturbed. In control embryos, there is a clear trend where more mature organs have higher expression of *pea3*, presumably because of longer or higher exposure to microluminal FGF. Thus, unperturbed first and third organs provided internal controls. Laser micropuncture causes the second organ to have significantly lower transcript counts than the less mature third organ, a result never observed in non-micropunctured controls.

The effect of micropuncture on FGF signalling was further investigated by performing smFISH protocol immediately after ( $t < 2$  min), 1 h after and 4 h after micropuncture. The transcript count per cell of the punctured organ 2 was normalized internally to the unperturbed organ 3.

To test the effect of microluminal FGF loss after micropuncture on collective cell behaviour, different organs before their depositions were micropunctured and the end-point spacing of the corresponding organ was compared with the unperturbed siblings. For Fgf3–GFP overexpression experiments, embryos were induced with 20  $\mu\text{M}$  inducer at 24 h.p.f. At 28 h.p.f., embryos with similar Fgf3–GFP overexpression levels were pre-selected to eliminate sample variability due to drug induction. Organ 2 or 3 was punctured while organ 1 or organ 2 was being deposited, respectively. For secGFP expression experiments, the same strategy was followed without pre-screening for expression levels as secGFP is a neutral marker to visualize intact luminal space.

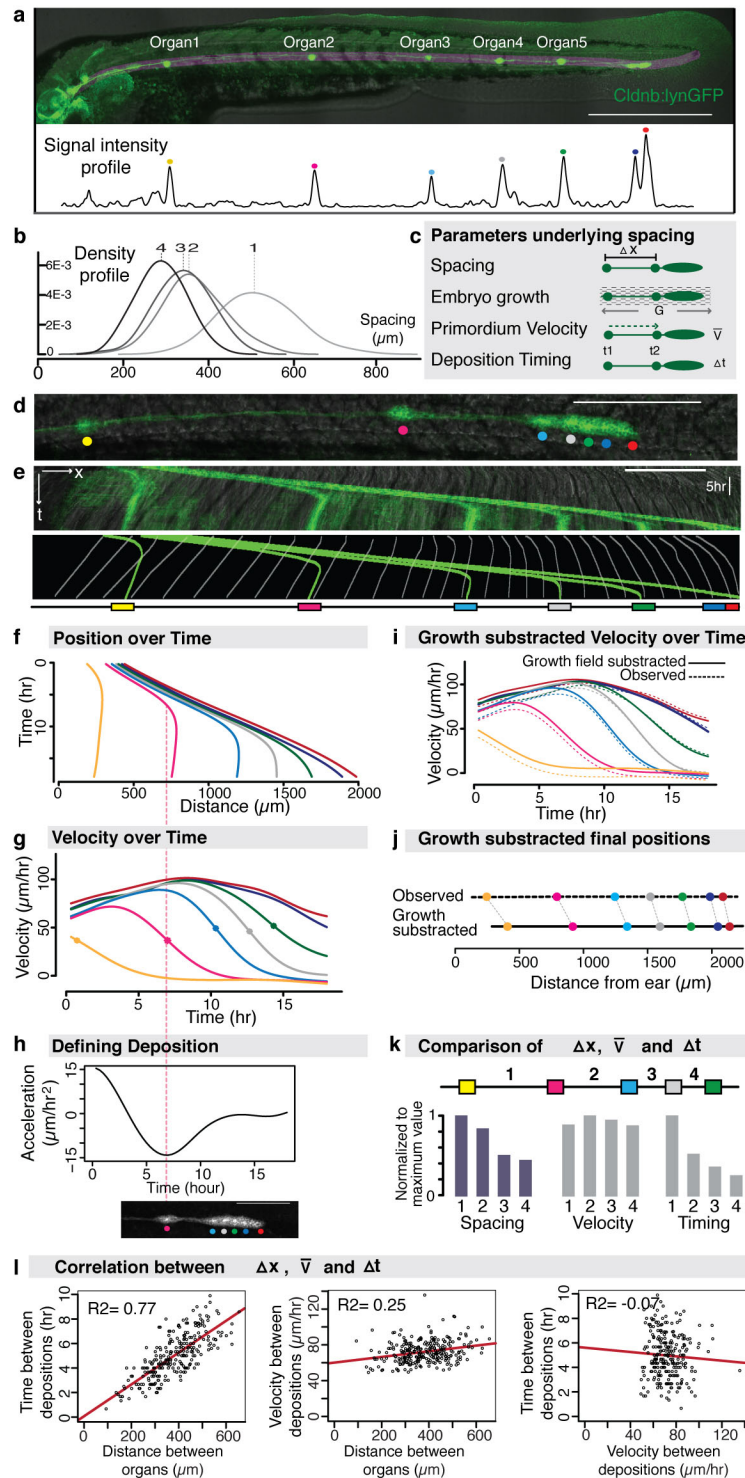
**Statistical analysis.** All statistical analysis used R. A non-parametric Wilcoxon rank-sum test (two sided) was used to compare two groups. Sample sizes ( $n$ ) and  $P$  values ( $P$ ) for each experiment are indicated in figure legends and exact  $P$  values are listed below. The statistical dependence between two variables was assessed with Spearman's rank correlation coefficient.

In experiments of organ patterning, more than 50 samples were acquired and automatically analysed to ensure adequate statistical power. For the experiments of transplantation, micropuncture and smFISH analysis, the number of samples was mainly constrained by the complexity of the experimental procedure, data acquisition and analysis capacity. In each experiment, data were analysed after the completion of data collection. To prevent selection bias, samples of stage and genotype matched pools were randomly divided into experimental groups. Additionally, data were analysed automatically where possible to avoid subjective assessments (for example, analysis of organ patterning and smFISH transcript counts).

Boxplots are standard box and whisker plots showing median and interquartile range. For all the experiments, original data points were displayed as scatter plots on top of boxplots using the beeswarm package<sup>42</sup>, allowing direct examination of the variance and distribution of the samples.

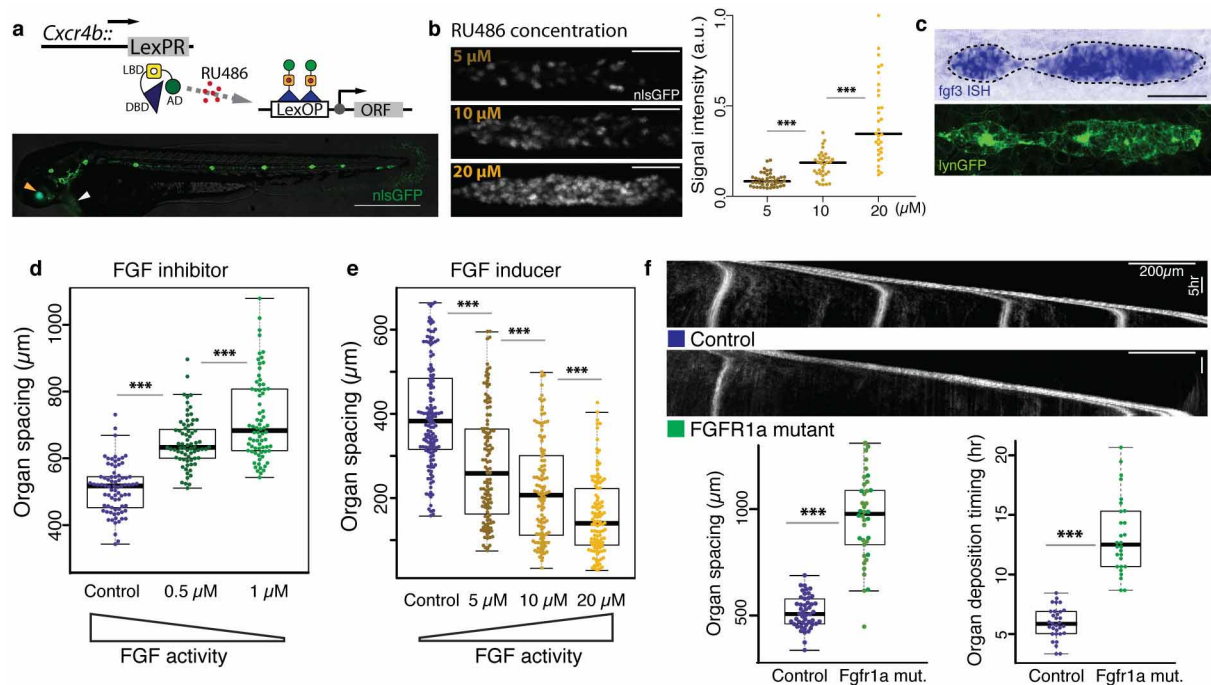
**$P$  values and sample sizes.** Fig. 1b: Spearman  $r^2 = 0.77$ ,  $N = 82$ ,  $n = 260$ . Fig. 1c:  $n_{\text{ctrl}} = 82$ ,  $n_{05} = 114$ ,  $n_1 = 104$ ,  $P_{\text{ctrl-05}} = 3.94 \times 10^{-24}$ ,  $P_{05-1} = 6.99 \times 10^{-29}$ . Fig. 1d:  $n_{\text{ctrl}} = 64$ ,  $n_5 = 57$ ,  $n_{10} = 65$ ,  $n_{20} = 58$ ,  $P_{\text{ctrl-5}} = 3.87 \times 10^{-5}$ ,  $P_{5-10} = 3.71 \times 10^{-4}$ ,  $P_{10-20} = 1.36 \times 10^{-5}$ . Fig. 3d:  $N_{\text{ctrl}} = 38$ ,  $N_{\text{mosaic}} = 114$ ,  $n_{\text{ctrl}} = 141$ ,  $n_{\text{neg}} = 170$ ,  $n_{\text{pos}} = 240$ ,  $P_{\text{ctrl-neg}} = 0.9203$ ,  $P_{\text{ctrl-pos}} = 1.29003 \times 10^{-38}$ ,  $P_{\text{neg-pos}} = 3.80489 \times 10^{-32}$ . Fig. 3e:  $N_{\text{WT}} = 4$ ,  $N_{\text{inducer}} = 4$ ,  $N_{\text{inhibitor}} = 4$ ,  $P_{\text{WT-inducer}} = 0.0285$ ,  $P_{\text{WT-inhibitor}} = 0.0285$ . Fig. 3f: Spearman  $r^2 = 0.85$ ,  $N = 18$ . Fig. 3g:  $n_{\text{ctrl}} = 4$ ,  $n_{\text{neg}} = 9$ ,  $n_{\text{pos}} = 7$ ,  $P_{\text{ctrl-neg}} = 0.6042$ ,  $P_{\text{neg-pos}} = 0.0001748$ . Fig. 4b:  $N_{\text{WT}} = 6$ ,  $N_{\text{shroomMO}} = 6$ ,  $P = 0.00216$ . Fig. 4e:  $N_{\text{ctrl}} = 6$ ,  $N_{\text{puncture}} = 16$ ,  $P = 0.0003217$ . Fig. 4f: second organ:  $N_{\text{ctrl}} = 16$ ,  $N_{\text{puncture}} = 23$ ,  $P = 1.443 \times 10^{-8}$ ; third organ:  $N_{\text{ctrl}} = 7$ ,  $N_{\text{puncture}} = 7$ ,  $P = 0.002331$ . Fig. 4g:  $N_{\text{closed}} = 6$ ,  $N_{\text{open}} = 6$ ,  $P = 0.002165$ . Extended Data Fig. 1: Spearman  $N = 82$ ,  $n = 260$ ,  $x-t r^2 = 0.77$ ,  $x-v r^2 = 0.25$ ,  $v-t r^2 = -0.07$ . Extended Data Fig. 2b:  $N_5 = 44$ ,  $N_{10} = 34$ ,  $N_{20} = 32$ ,  $P_{5-10} = 2.53 \times 10^{-8}$ ,  $P_{10-20} = 4.03 \times 10^{-8}$ . Extended Data Fig. 2d:  $n_{\text{ctrl}} = 78$ ,  $n_{05} = 71$ ,  $n_1 = 74$ ,  $P_{\text{ctrl-05}} = 6.22 \times 10^{-14}$ ,  $P_{05-1} = 7.26 \times 10^{-4}$ . Extended Data Fig. 2e:  $n_{\text{ctrl}} = 109$ ,  $n_5 = 112$ ,  $n_{10} = 119$ ,  $n_{20} = 137$ ,  $P_{\text{ctrl-5}} = 1.33 \times 10^{-10}$ ,  $P_{5-10} = 7.06 \times 10^{-4}$ ,  $P_{10-20} = 2.46 \times 10^{-4}$ . Extended Data Fig. 2f: spacing:  $N_{\text{ctrl}} = 49$ ,  $N_{\text{mut}} = 39$ ,  $P = 7.69 \times 10^{-14}$ ; timing:  $N_{\text{ctrl}} = 31$ ,  $N_{\text{mut}} = 28$ ,  $P = 4.64 \times 10^{-11}$ . Extended Data Fig. 6d:  $N_{\text{WT}} = 9$ ,  $N_{\text{rescue}} = 13$ ,  $N_{\text{FGFmo}} = 14$ ,  $P_{\text{WT-rescue}} = 0.09$ ,  $P_{\text{rescue-FGFmo}} = 1.751 \times 10^{-6}$ . Extended Data Fig. 7b:  $N_{\text{control}} = 7$ ,  $N_{\text{transplants}} = 8$ ,  $n_{\text{control}} = 25$ ,  $n_{\text{neg}} = 17$ ,  $n_{\text{pos}} = 13$ ; spacing:  $P_{\text{ctrl-neg}} = 0.24$ ,  $P_{\text{ctrl-pos}} = 1.43 \times 10^{-5}$ ,  $P_{\text{neg-pos}} = 4.40 \times 10^{-5}$ ; timing:  $P_{\text{ctrl-neg}} = 0.07$ ,  $P_{\text{ctrl-pos}} = 1.23 \times 10^{-6}$ ,  $P_{\text{neg-pos}} = 4.09 \times 10^{-5}$ . Extended Data Fig. 9d:  $N_{0\text{h puncture}} = 6$ ,  $N_{0\text{h control}} = 5$ ,  $N_{1\text{h puncture}} = 6$ ,  $N_{1\text{h control}} = 5$ ,  $N_{4\text{h puncture}} = 6$ ,  $N_{4\text{h control}} = 5$ ,  $P_{0\text{h}} = 0.7922$ ,  $P_{1\text{h}} = 0.0043$ ,  $P_{4\text{h}} = 0.4286$ . Extended Data Fig. 9e:  $N_{\text{ctrl second organ}} = 22$ ,  $N_{\text{puncture second organ}} = 23$ ,  $N_{\text{ctrl third organ}} = 8$ ,  $N_{\text{puncture third organ}} = 9$ ,  $P_{\text{second organ}} = 6.928 \times 10^{-6}$ ,  $P_{\text{third organ}} = 0.0061$ .

31. Westerfield, M. *The Zebrafish Book* 5th edn (Univ. Oregon Press, 2007).
32. Haas, P. & Gilmour, D. Chemokine signaling mediates self-organizing tissue migration in the zebrafish lateral line. *Dev. Cell* **10**, 673–680 (2006).
33. Kwan, K. M. et al. The Tol2kit: a multisite gateway-based construction kit for Tol2 transposon transgenesis constructs. *Dev. Dyn.* **236**, 3088–3099 (2007).
34. Kwon, H. J. & Riley, B. B. Mesendodermal signals required for otic induction: Bmp-antagonists cooperate with Fgf and can facilitate formation of ectopic otic tissue. *Dev. Dyn.* **238**, 1582–1594 (2009).
35. Norton, W. H. J., Ledin, J., Grandel, H. & Neumann, C. J. HSPG synthesis by zebrafish Ext2 and Extl3 is required for Fgf10 signalling during limb development. *Development* **132**, 4963–4973 (2005).
36. Preibisch, S., Saalfeld, S. & Tomancak, P. Globally optimal stitching of tiled 3D microscopic image acquisitions. *Bioinformatics* **1**, 1463–1465 (2009).
37. Morhac, M. Peaks: Peaks. R package v.0.2. (<http://cran.r-project.org/web/packages/Peaks/Peaks.pdf>, 2012).
38. Wand, M. KernSmooth: functions for kernel smoothing for Wand & Jones (1995). R package v.2.23-10. (<http://cran.r-project.org/web/packages/KernSmooth/KernSmooth.pdf>, 2011).
39. Kremer, J. R., Mastrorade, D. N. & McIntosh, J. R. Computer visualization of three-dimensional image data using IMOD. *J. Struct. Biol.* **116**, 71–76 (1996).
40. Rapsomaniki, M. A. et al. easyFRAP: an interactive, easy-to-use tool for qualitative and quantitative analysis of FRAP data. *Bioinformatics* **1**, 1800–1801 (2012).
41. Raj, A., van den Bogaard, P., Rifkin, S. A., van Oudenaarden, A. & Tyagi, S. Imaging individual mRNA molecules using multiple singly labeled probes. *Nature Methods* **5**, 877–879 (2008).
42. Eklund, A. beeswarm: The bee swarm plot, an alternative to stripchart. R package v.0.1.6 (<http://cran.r-project.org/web/packages/beeswarm/beeswarm.pdf>, 2013).



**Extended Data Figure 1 | Quantitative analysis of lateral line organ deposition.** **a**, Posterior lateral line organs at 2 d.p.f. (*cldnb:lynGFP*). Organ positions were identified from intensity profiles using peakFinder\_R. **b**, Density profile of distance between consecutive organ positions (first, second, third and fourth spacing interval; see Fig. 1a). **c**, List of potential parameters affecting organ spacing. **d**, *cldnb:lynGFP* and brightfield overlay image. Spheres indicate colour code representing individual organs used in further analysis. **e**, Upper panel, kymograph ( $x-t$  graph) from a 17.6 h time-lapse movie, where the  $y$  axis represents time and the  $x$  axis represents distance. Lower panel, segmented kymograph of primordium migration (green) and myotome growth (dashed lines) through time. **f**, **g**, Calculated position (**f**) and velocity (**g**) of each organ

through time. Asterisk shows the time point when organ disengages from the migrating collective. **h**, Second organ acceleration through time. Organ deposition is defined as the time where acceleration is minimum. **i**, Growth-effect-subtracted velocity of each organ through time (solid lines) versus observed velocities (dashed line). **j**, Reconstruction of organ positions from growth-subtracted velocities. **k**, Comparing spacing, average velocity and time between consecutive depositions for first, second, third and fourth interval (normalized to maximum). **l**, Correlation of time, distance and average velocities between consecutive depositions. Statistics: Spearman  $N = 82$ ,  $n = 260$ ,  $x-t$   $r^2 = 0.77$ ,  $x-v$   $r^2 = 0.25$ ,  $v-t$   $r^2 = -0.07$ . Scale bars, 500  $\mu\text{m}$  (**a**), 200  $\mu\text{m}$ , 5 h (**d**, **e**).

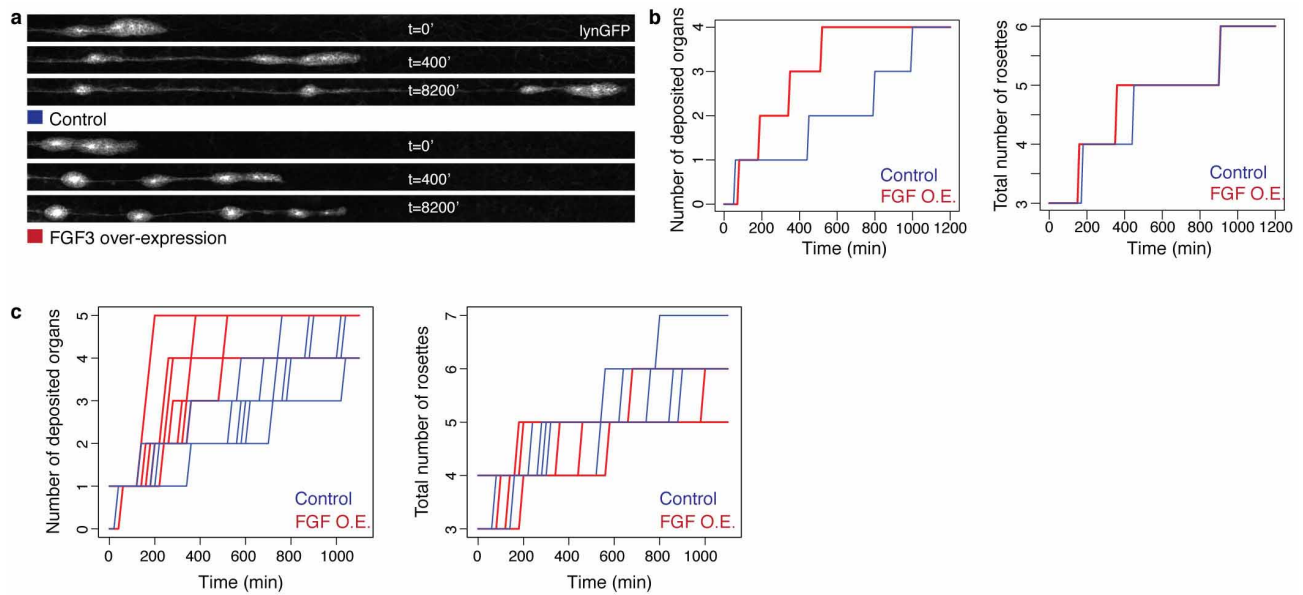


**Extended Data Figure 2 | 'Tunable' drug-inducible gene expression with LexPR and quantification of dose-dependent response to FGF signalling.**

**a**, Schema shows transactivator LexPR expressed under the control of *Cxcr4b* promoter (*Cxcr4b:LexPR*) driving expression of LexOP-coupled coding sequences upon addition of inducer RU486 (above). Image of *Cxcr4b:LexPR*-driven expression of *lexOP:nlsGFP* showing spatially restricted expression upon RU486 treatment. Scale bar, 500  $\mu\text{m}$  (*cry:eCFP* 'crystal eye' marker: orange arrow; *clmc2:GFP* 'bleeding heart' marker: white arrow). **b**, Mean fluorescence intensity projection of *Cxcr4b:LexPR, LexOP:nlsGFP* primordium treated with 5 ( $n = 44$ ), 10 ( $n = 34$ ) and 20  $\mu\text{M}$  ( $n = 32$ ) RU486. Scale bar, 50  $\mu\text{m}$ . Plot shows quantification of signal intensity after 4 h of RU486

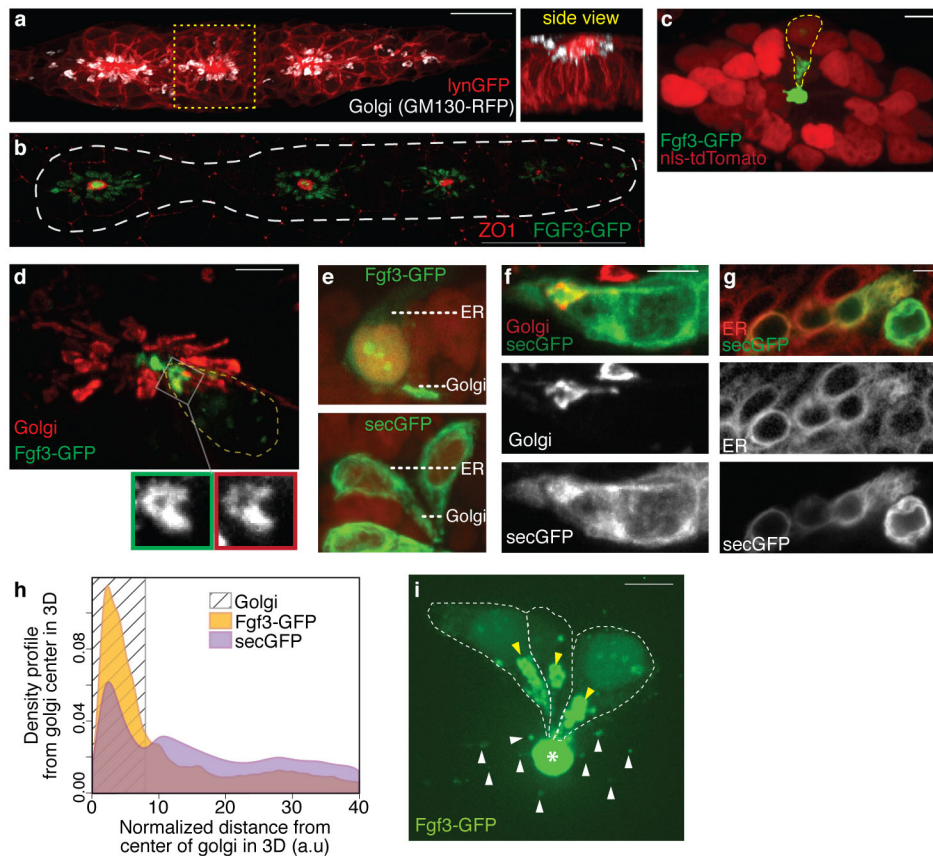
induction ( $P_{5-10} = 2.53 \times 10^{-8}$ ,  $P_{10-20} = 4.03 \times 10^{-8}$ ). **c**, Colorimetric *in situ* hybridization of *fgf3* mRNA in *Cxcr4b:LexPR, LexOP:Fgf3-GFP* showing uniform expression. Scale bar, 50  $\mu\text{m}$ . **d**, **e**, Organ spacing in FGF inhibitor- and inducer-treated embryos at 2 d.p.f. **d**, Quantification of organ spacing ( $n = 78, 71, 74$ ,  $P_{\text{ctrl-0.5}} = 6.22 \times 10^{-14}$ ,  $P_{0.5-1} = 7.26 \times 10^{-4}$ ) in SU5402-treated samples. **e**, Quantification of organ spacing ( $n = 109, 112, 119, 137$ ,  $P_{\text{ctrl-5}} = 1.33 \times 10^{-10}$ ,  $P_{5-10} = 7.06 \times 10^{-4}$ ,  $P_{10-20} = 2.46 \times 10^{-4}$ ) in RU486-treated samples. **f**, Organ depositions in WT and homozygous *fgfr1a*<sup>13R705H</sup> mutants shown by kymographs of 21 h time-lapse movies. Quantification of spacing ( $n = 49, 39$ ,  $P = 7.69 \times 10^{-14}$ ) and deposition timing ( $n = 31, 28$ ,  $P = 4.64 \times 10^{-11}$ ) between organs (first interval). Scale bar, 200  $\mu\text{m}$ , 5 h.





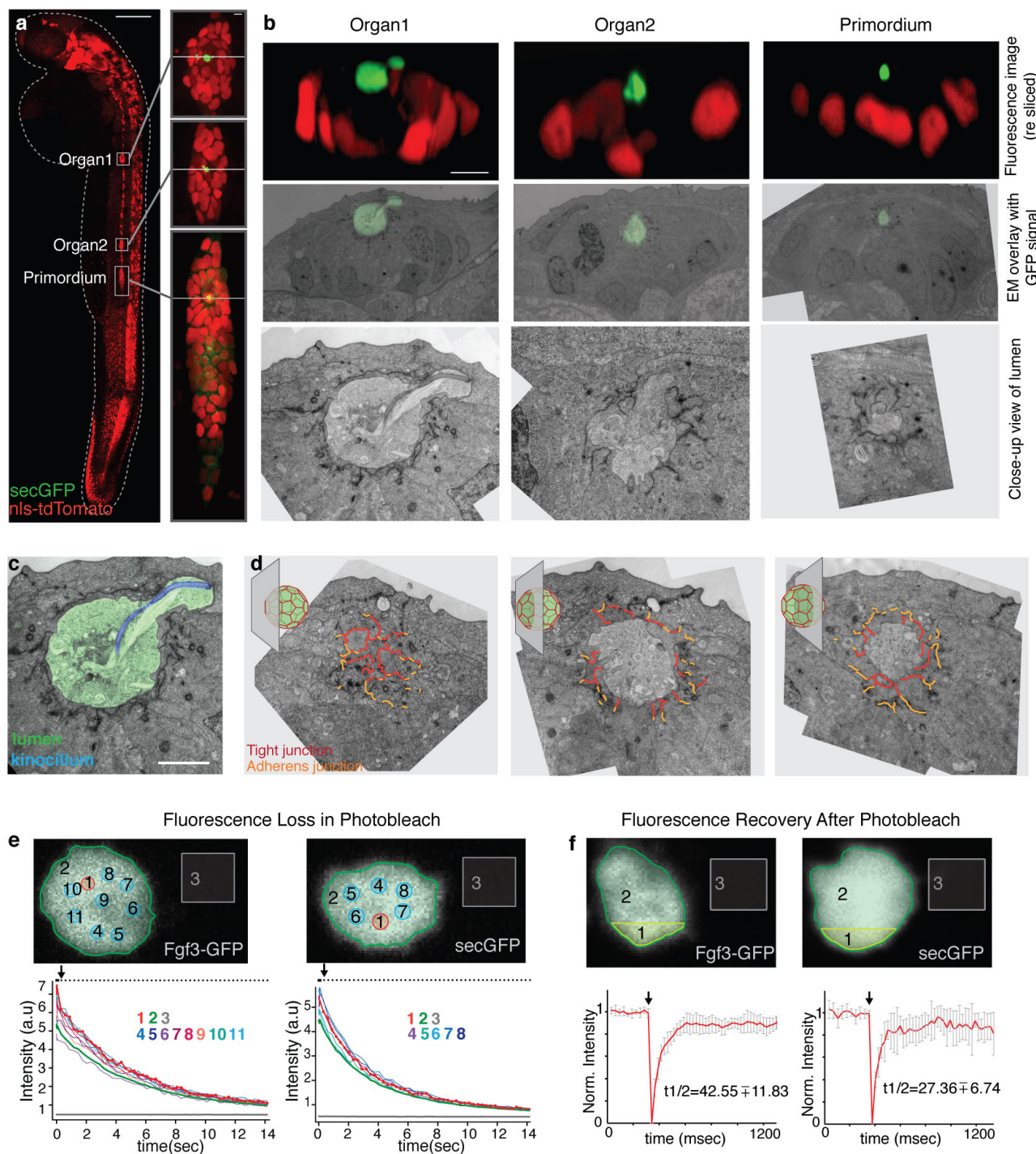
**Extended Data Figure 3 | Organ deposition and rosette formation rate upon Fgf3-GFP overexpression.** **a**, *cldnb:lynGFP* embryos showing comparison of organ deposition and rosette formation rate upon *lexOP:fgf3*-GFP overexpression. **b**, **c**, Comparisons of total number of organs deposited

(left) and total number of organ progenitor rosettes assembled (right) through time in control (blue) and *lexOP:fgf3*-GFP (red) embryos. Only organ deposition timing shows a clear difference between these conditions. **c**, Plots showing multiple examples of data in **b** ( $n = 7, 7$ ).



**Extended Data Figure 4 | SecGFP and Fgf3-GFP localization in apically polarized secretory path.** **a**, Golgi, labelled by GM130-tdTomato (white) mRNA injection, are localized apically around rosette centres in lateral line primordium (*cldnb:lynGFP*, red). Scale bar, 20  $\mu$ m. **b**, Maximum projection of apical optical sections of a transgenic *lexOP:fgf3-GFP* primordium, counterstained for ZO1, shows intracellular Fgf3-GFP signal around rosette centres in addition to luminal signal. Scale bar, 50  $\mu$ m. **c**, Single cell expressing Fgf3-GFP feeds the central microlumen through apical secretion (expressing cell indicated with yellow dashed line). Scale bar, 5  $\mu$ m. **d**, Mosaic primordium showing apically localized intracellular Fgf3-GFP signal co-localizes with Golgi marker GM130-tdTomato. Scale bar, 5  $\mu$ m. **e**, Intracellular Fgf3-GFP and

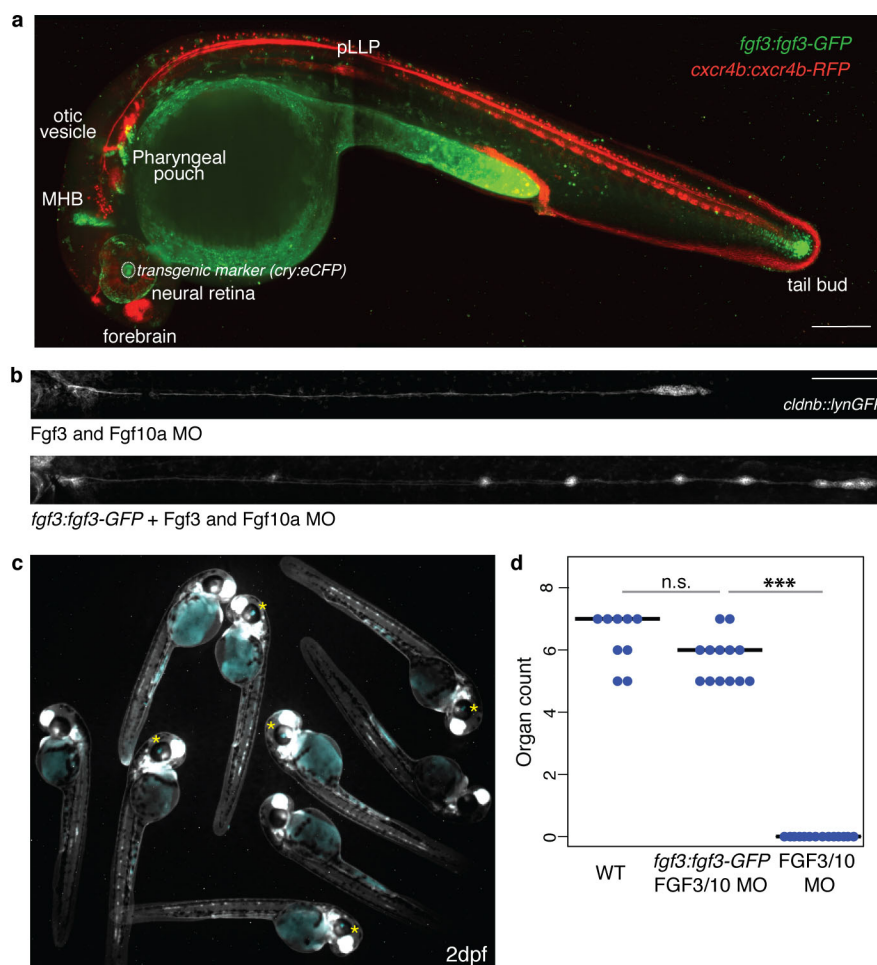
secGFP localization at secretory path. **f**, Golgi (GM130-tdTomato) co-labelling with secGFP. Scale bar, 5  $\mu$ m. **g**, Endoplasmic reticulum (mKate2-KDEL) co-labelling with secGFP. Scale bar, 5  $\mu$ m. **h**, Signal distribution of secGFP and Fgf3-GFP in three dimensions within the expressing cell where Golgi was taken as a central point. Comparison of Fgf3-GFP and secGFP density profiles suggests that Fgf3-GFP is more pronounced in Golgi ( $n_{\text{secGFP}} = 5$ ,  $n_{\text{Fgf3-GFP}} = 4$ ). **i**, Imaging of Fgf3-GFP-expressing clones (white dashed lines) with high sensitivity reveals Golgi localization (yellow arrowheads) of Fgf3-GFP in expressing cells close to the microlumen (asterisk) and intracellular vesicles in connected non-expressing cells (white arrowheads). No extracellular signal besides microluminal accumulation was detected. Scale bar, 5  $\mu$ m.



**Extended Data Figure 5 | CLEM analysis of microlumen structure and FLIP/FRAP analysis of microluminal pools.** **a**, Overview of *lexOP:secGFP; cxcr4b:nls-tdTomato* embryo used for CLEM; two organs and migrating primordium were targeted for further processing. Scale bar, 200  $\mu\text{m}$ . **b**, Re-sliced middle section of targeted organ centres, overlay of secGFP signal with corresponding EM slice (scale bar, 5  $\mu\text{m}$ ) and close-up view of microlumina. **c**, Close-up view of luminal cavity (green) distorted by kinocilium (blue). **d**, Traced tight junctions (red) and adherens junctions (orange) at three

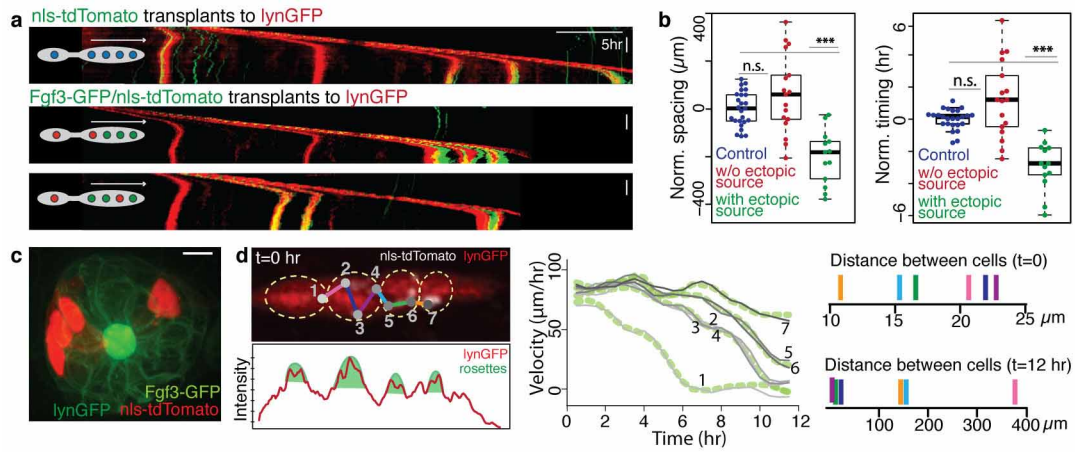
cross-sections of microlumen. **e**, Setup of FLIP experiment on Fgf3-GFP and secGFP pool highlighting repetitively bleached region (0.73  $\mu\text{m}$  diameter, red circle) and regions used for total pool (green circle), background (grey box) and readout (blue circles) measurements. Plots show mean intensity of described ROIs over time. **f**, FRAP experiment on secGFP and Fgf3-GFP pools with a strip ROI. Mean normalized recovery curves (mean  $\pm$  s.d.,  $N = 7$ ) and calculated half time of recovery. Arrow indicates start of bleaching.





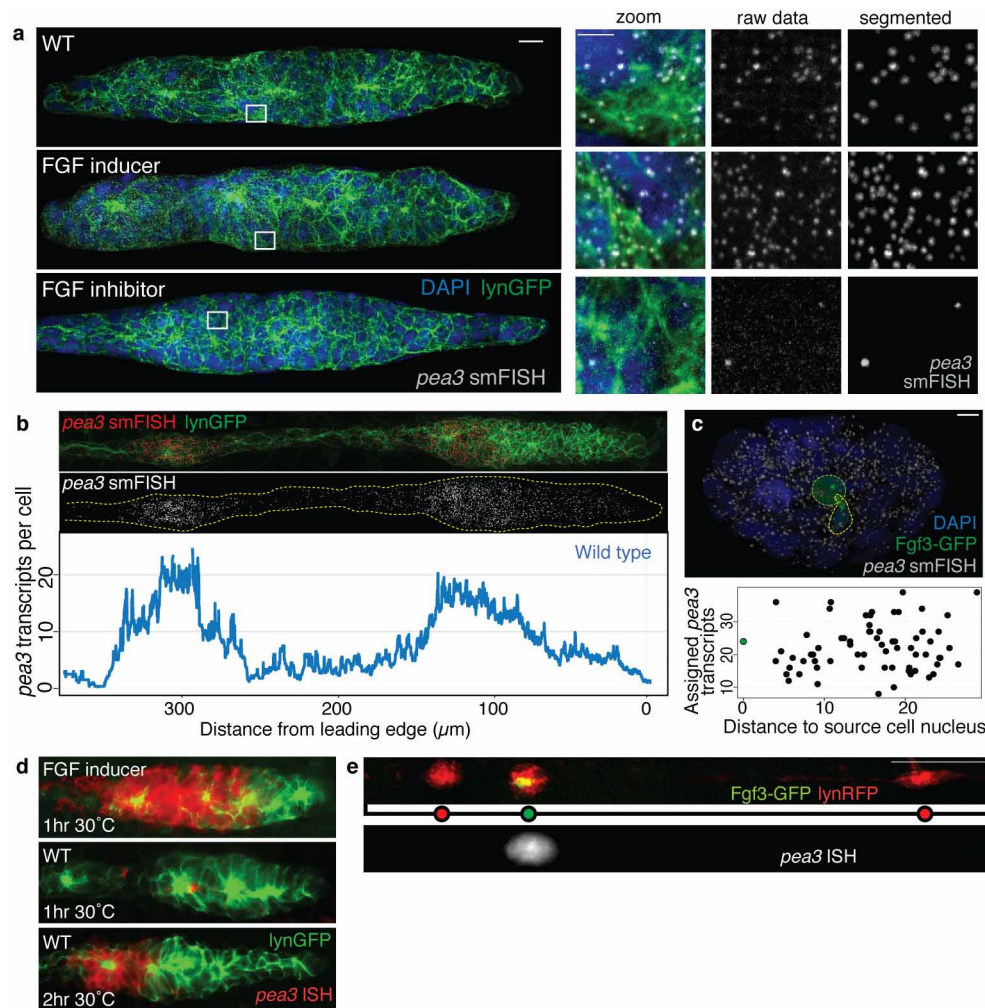
**Extended Data Figure 6 | BAC *fgf3:fgf3-GFP* rescues FGF loss of function in lateral line.** **a**, BAC *fgf3:fgf3-GFP* line showing expression in known Fgf3 expression domains (28 h.p.f.). Scale bar, 200  $\mu$ m. **b**, Loss-of-organ deposition phenotype Fgf3/10a morphant embryos (Fgf3/10a MO, upper) is rescued by BAC *fgf3:fgf3-GFP* transgene (lower). **c**, Low-magnification image showing Fgf3/10 morphants, with BAC *fgf3:fgf3-GFP* rescued siblings, distinguished

by crystal eye transgene marker (yellow star). Scale bar, 200  $\mu$ m (**b**). **d**, Quantification of rescue by comparing organ counts of WT, *fgf3:fgf3-GFP* with Fgf3/10 MO background and Fgf3/10a MO alone at 2 d.p.f. ( $N_{WT} = 9$ ,  $N_{rescue} = 13$ ,  $N_{Fgf3/10a\_MO} = 14$ ,  $P_{WT-rescue} = 0.09$ ,  $P_{rescue-Fgf3/10a\_MO} = 1.751 \times 10^{-6}$ ).



**Extended Data Figure 7 | FGF signalling range is restricted to individual organ progenitors.** **a**, Kymographs of mosaic Fgf3-GFP expression generated via cell transplantation. *lexOP:fgf3-GFP/cxcr4b:nls-tdTomato*-expressing clones (green) in the *cldnb:lynGFP* line (red) cause rapid arrest of migration. The phenotype only becomes apparent when the organ reaches tissue rear. (Colour code: organs with ectopic FGF source in green; organs without ectopic FGF source in red; organs of control transplants in blue.) Scale bars, 200  $\mu\text{m}$ , 5 h. **b**, Quantification of spacing and deposition timing of organs from mosaic Fgf3-GFP transplants, normalized by mean values of control embryos for each interval ( $N_{\text{control}} = 7$ ,  $N_{\text{transplants}} = 8$ ,  $n_{\text{control}} = 25$ ,  $n_{\text{neg}} = 17$ ,  $n_{\text{pos}} = 13$ ; spacing:  $P_{\text{ctrl-neg}} = 0.24$ ,  $P_{\text{ctrl-pos}} = 1.43 \times 10^{-5}$ ,  $P_{\text{neg-pos}} = 4.40 \times 10^{-5}$ ; timing:  $P_{\text{ctrl-neg}} = 0.07$ ,  $P_{\text{ctrl-pos}} = 1.23 \times 10^{-6}$ ,  $P_{\text{neg-pos}} = 4.09 \times 10^{-5}$ ).

**c**, Close-up view of Fgf3-GFP (green)/nls-tdTomato- (red) expressing clones in *cldnb:lynGFP*- (green) expressing organ, showing cells in different positions feed in the central microlumen. Scale bar, 5  $\mu\text{m}$ . **d**, Tracking of WT transplanted cells (nuclei marked with grey dots and numbered) relative to organ centres in *cldnb:lynGFP* primordium (red). Yellow circles represent each organ unit. Middle panel: calculated velocities for each tracked nucleus (grey lines) and organ centres (green lines) reveal that migration of individual cells is in synchrony with the belonged organ unit independent of their position. Right panel: distance between consecutive tracked cells at the beginning and end of the time-lapse movie shows that initial distance is not a reliable indicator of final cell positions.

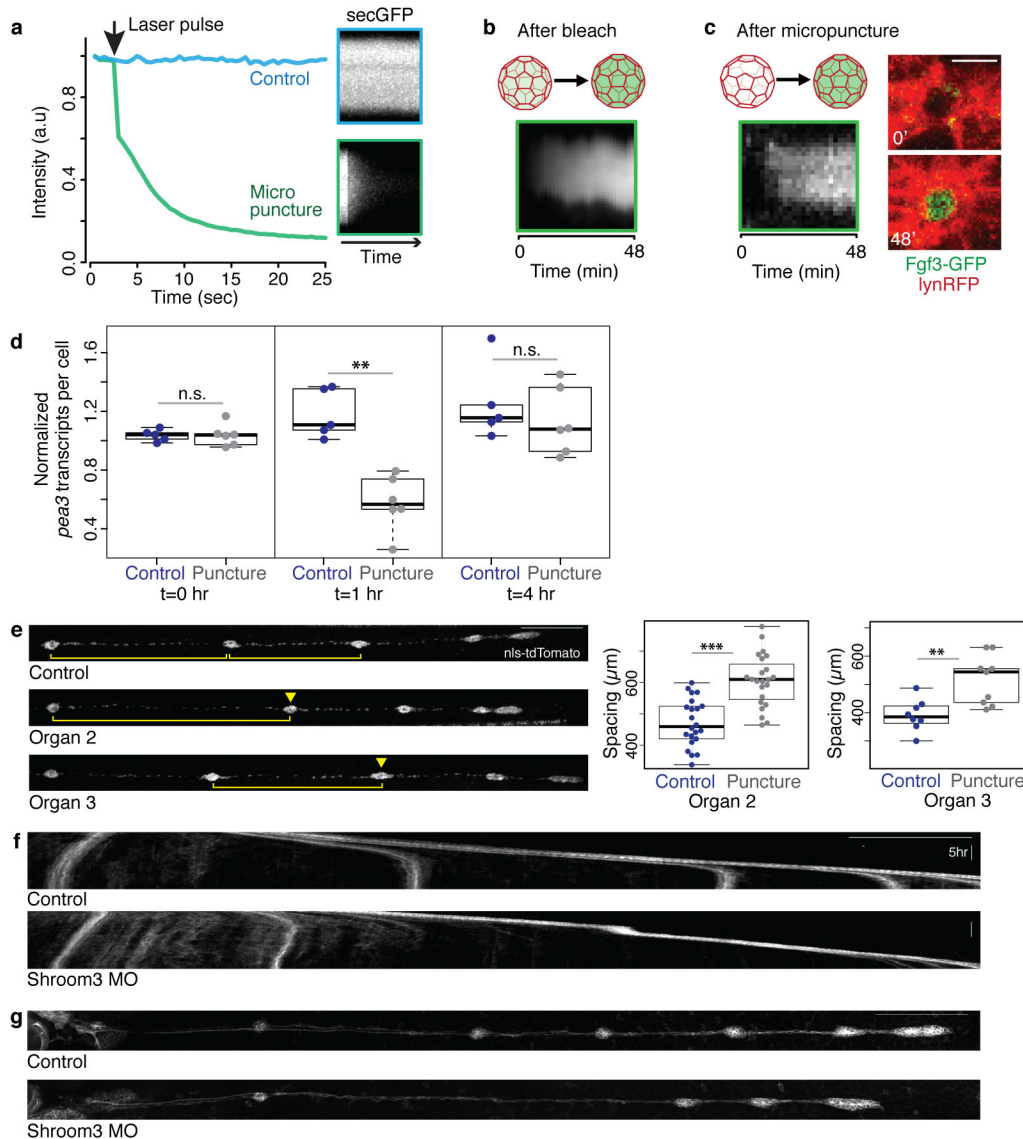


#### Extended Data Figure 8 | smFISH analysis of FGF target-gene regulation.

**a**, *Pea3* smFISH on WT, 15  $\mu\text{M}$  FGF inducer- and 4  $\mu\text{M}$  FGF inhibitor-treated primordia (*cldnb:lynGFP* in green, DAPI staining in blue, *pea3* mRNAs in white). Scale bar, 5  $\mu\text{m}$ . Close-up view of the dashed boxes shown as raw image (middle) and segmented *pea3* transcripts (right). Scale bar, 2  $\mu\text{m}$ . **b**, Image of *pea3* smFISH in WT primordium (above); profile plot shows *pea3* transcripts per cell over distance from leading edge (below). **c**, *Pea3* smFISH in an organ with single Fgf3-GFP-expressing cell. Number of *pea3* transcripts

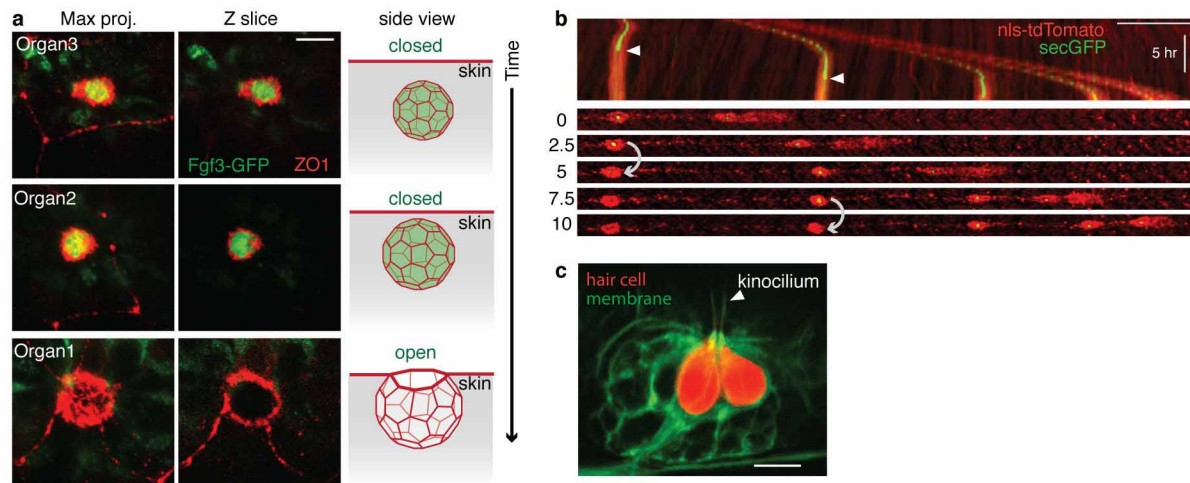
assigned to each nucleus does not show increase towards the expressing cell. Scale bar, 5  $\mu\text{m}$ . **d**, Colorimetric *in situ* hybridization of *pea3* mRNA showing high expression levels upon Fgf3-GFP induction, visible after 1 h colour reaction (30  $^{\circ}\text{C}$ ), whereas expression in WT is hardly detectable. However, increasing reaction time reveals *pea3* mRNA signal in WT primordia. **e**, Colorimetric *in situ* hybridization (30  $^{\circ}\text{C}$ , 0.5 h) of *pea3* RNA in mosaic Fgf3-GFP expression shows detectable *pea3* only in the expressing organ. Scale bar, 100  $\mu\text{m}$ .





**Extended Data Figure 9 | Characterization of luminal integrity and function upon mechanical and genetic perturbation.** **a**, Plot of secGFP pool fluorescence intensity upon micropuncture (green). Kymographs show the time-lapse imaging of the secGFP pool used for the plot. **b, c**, Luminal Fgf3-GFP signal recovery of whole pool bleached (left) and micro-punctured (right) organs during 48 min of acquisition. Kymographs show time-lapse imaging of Fgf3-GFP pool. Single time points of time-lapse imaging after micro-puncture (right). Scale bar, 5  $\mu$ m. **d**, Quantification of *pea3* transcript levels at  $t = 0$  h, 1 h and 4 h after micropuncture of organ 2 expressing *lexOP:fgf3-GFP*. Unperturbed organ 3 was used for normalization. Comparison of control and punctured organs suggests that *pea3* levels are

normal immediately after puncture, are reduced 1 h later and recovered by 4 h ( $N_{0\text{ h puncture}} = 6$ ,  $N_{0\text{ h control}} = 5$ ,  $N_{1\text{ h puncture}} = 6$ ,  $N_{1\text{ h control}} = 5$ ,  $N_{4\text{ h puncture}} = 6$ ,  $N_{4\text{ h control}} = 5$ ,  $P_{0\text{ h}} = 0.7922$ ,  $P_{1\text{ h}} = 0.0043$ ,  $P_{4\text{ h}} = 0.4286$ ). **e**, Organ deposition delay upon lumina micropuncture of secGFP-expressing second and third organs ( $N_{\text{ctrl second organ}} = 22$ ,  $N_{\text{puncture second organ}} = 23$ ,  $N_{\text{ctrl third organ}} = 8$ ,  $N_{\text{puncture third organ}} = 9$ ,  $P_{\text{second organ}} = 6.928 \times 10^{-6}$ ,  $P_{\text{third organ}} = 0.0061$ ). Scale bar, 200  $\mu$ m. **f, g**, Shroom3 morphant primordia show intervals with no or delayed deposition. **f**, Kymographs of shroom3 MO and control. Scale bars, 200  $\mu$ m, 5 h. **g**, Organ pattern in shroom3 MO and control at 2 d.p.f. Scale bar, 200  $\mu$ m.



**Extended Data Figure 10 | Loss of microlumen pool upon fusion to overlying skin.** **a**, Microlumen of maturing organs fuses with the skin and the diffusible content (Fgf3-GFP in green) disappears. Tight junctions marking microlumen and skin borders are revealed by ZO1 immunofluorescence (red). Cartoon displaying the sequence of events (right). Scale bar, 5  $\mu$ m. **b**, Kymograph and single time-points from time-lapse imaging of secGFP,

nls-tdTomato-expressing embryo. SecGFP signal disappears as microlumen opens (arrowheads in kymograph show opening of microlumina). Scale bars, 200  $\mu$ m, 5 h. **c**, Side view of a maturing organ with kinocilia protruding out of the organ (*cldnb:lynGFP* in green, central cell *atoh1a:tdTomato* in red). Scale bar, 5  $\mu$ m.

# PLETHORA gradient formation mechanism separates auxin responses

Ari Pekka Mähönen<sup>1,2,3\*</sup>, Kirsten ten Tusscher<sup>4\*</sup>, Riccardo Siligato<sup>1,3</sup>, Ondřej Smetana<sup>1,3</sup>, Sara Díaz-Triviño<sup>2,5</sup>, Jarkko Salojärvi<sup>3</sup>, Guy Wachsman<sup>2</sup>, Kalika Prasad<sup>2</sup>, Renze Heidstra<sup>2,5</sup> & Ben Scheres<sup>2,5</sup>

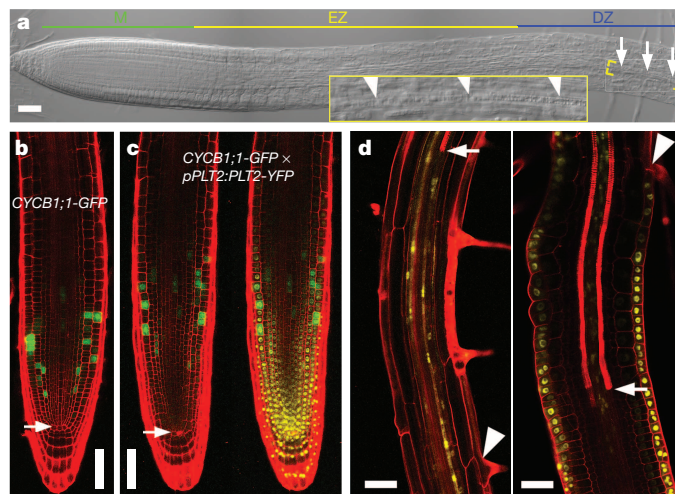
During plant growth, dividing cells in meristems must coordinate transitions from division to expansion and differentiation, thus generating three distinct developmental zones: the meristem, elongation zone and differentiation zone<sup>1</sup>. Simultaneously, plants display tropisms, rapid adjustments of their direction of growth to adapt to environmental conditions. It is unclear how stable zonation is maintained during transient adjustments in growth direction. In *Arabidopsis* roots, many aspects of zonation are controlled by the phytohormone auxin and auxin-induced PLETHORA (PLT) transcription factors, both of which display a graded distribution with a maximum near the root tip<sup>2–12</sup>. In addition, auxin is also pivotal for tropic responses<sup>13,14</sup>. Here, using an iterative experimental and computational approach, we show how an interplay between auxin and PLTs controls zonation and gravitropism. We find that the PLT gradient is not a direct, proportionate readout of the auxin gradient. Rather, prolonged high auxin levels generate a narrow PLT transcription domain from which a gradient of PLT protein is subsequently generated through slow growth dilution and cell-to-cell movement. The resulting PLT levels define the location of developmental zones. In addition to slowly promoting PLT transcription, auxin also rapidly influences division, expansion and differentiation rates. We demonstrate how this specific regulatory design in which auxin cooperates with PLTs through different mechanisms and on different timescales enables both the fast tropic environmental responses and stable zonation dynamics necessary for coordinated cell differentiation.

We have previously shown that four PLT transcription factors with graded distribution (PLT1, PLT2, PLT3 and BBM (also known as PLT4)) are necessary for stem cell maintenance and cell division in the root<sup>8,9</sup>. Furthermore, correlation of PLT protein levels with the developmental transitions that define root zonation (Fig. 1a) suggests a dosage-dependent control by PLTs<sup>9</sup>. However, two issues remain unresolved.

First, the precise relationship between PLT dosage and the location and size of the stem cell domain has not been established. Therefore, we investigated whether different PLT levels mediate the distinction between slowly dividing stem cells and fast dividing transit amplifying cells. The addition of extra copies of PLT2 led to an enlarged meristem and shootward shift of the high-division-rate domain (Fig. 1b, c and Extended Data Fig. 1a, b), indicating that the highest dose of PLT2 slows down division rates as observed in the stem cell niche, while medium levels trigger high division rates shootward from the stem cell region.

Second, it remained to be established whether, similar to stem cell factors in the animal kingdom, PLT transcription factors repress differentiation. In that case, expression of PLT2 in one cell type should be sufficient to block differentiation locally while allowing differentiation of other cell types. To test this, we induced yellow fluorescent protein (YFP)-tagged PLT2 using either a protoxylem and the associated pericycle-specific promoter *pAHP6* (ref. 15) or an epidermal/lateral root cap promoter *pWER*<sup>16</sup>. *pAHP6:XVE>>PLT2-YFP* induction inhibited protoxylem differentiation

and caused local ectopic cell proliferation while root hair differentiation proceeded normally. Reciprocally, *pWER:XVE>>PLT2-YFP* induction triggered local inhibition of root hair differentiation and ectopic cell division while protoxylem differentiation proceeded normally (Fig. 1d and Extended Data Fig. 1c). Furthermore, induction of PLT2 inhibited cell expansion, which is generally considered to be an early step in cell differentiation. The speed at which PLTs control expansion suggests that the decline in PLT levels along the gradient determines the transition to differentiation (Supplementary Notes and Extended Data Fig. 1d, e). Finally, we tested whether this differentiation threshold was imposed also by physiologically relevant PLT concentrations. Reduction of PLT2 by inducible RNA interference (RNAi) in the *plt1,3,4* mutant, which solely depends on PLT2 to form functional meristems<sup>9</sup>, indeed triggered meristem cell expansion and differentiation (Extended Data Fig. 1f). Taken together, our results show that the PLT protein gradient shape defines the location of at least two boundaries: the boundary between slowly and rapidly cycling cells, and the shootward boundary of the meristem.



**Figure 1 | PLT levels define zonation boundaries.** **a**, Zonation of 4-day-old wild-type root. Arrows and arrowheads indicate youngest protoxylem cell. The meristem (M), expansion (EZ) and differentiation (DZ) zones are highlighted. **b**, **c**, Frequent cell division, monitored by the G2/M-phase cell cycle marker CYCB1:1-GFP, occurs close to the quiescent centre (arrow) in wild-type meristem (**b**). This domain shifts shootward with increased PLT2 dosage (that is, homozygote *pPLT2:PLT2-YFP* in Col background; green and green/yellow channels shown) (**c**). **d**, Twenty-four hours induction of PLT2-YFP in the vascular tissue (left) locally inhibits xylem differentiation (arrow, first xylem element), while PLT2-YFP induction in epidermis (right) inhibits root hair formation (arrowhead, first root hair). Propidium iodide highlights cell wall and protoxylem in **b–d**. Scale bars, 50  $\mu$ m.

<sup>1</sup>Institute of Biotechnology, University of Helsinki, Helsinki 00014, Finland. <sup>2</sup>Molecular Genetics, Department of Biology, Utrecht University, Utrecht 3584 CH, the Netherlands. <sup>3</sup>Department of Biosciences, University of Helsinki, Helsinki 00014, Finland. <sup>4</sup>Theoretical Biology and Bioinformatics, Utrecht University, Utrecht 3584 CH, the Netherlands. <sup>5</sup>Plant Developmental Biology, Wageningen University Research, Wageningen 6708 PB, the Netherlands.

\*These authors contributed equally to this work.



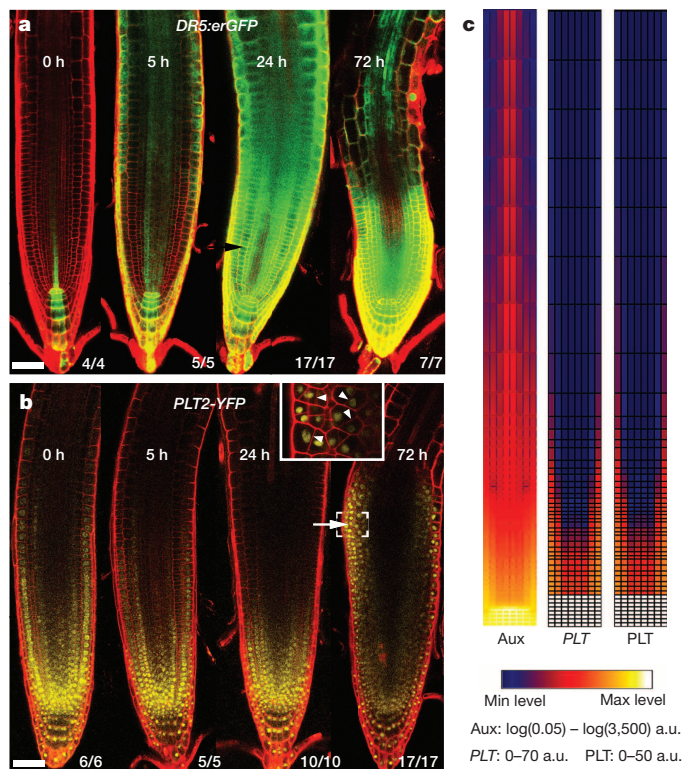
PLT gradients have been considered to be generated at the transcriptional level, based on the similarity of transcriptional and translational PLT–fluorescent protein fusion gradients<sup>9</sup>. *PLT* transcription requires auxin response factors (ARFs)<sup>8,12</sup>, and since auxin is also present in a graded pattern<sup>4,5</sup>, it was postulated that the PLT gradient may be a readout of the auxin gradient. To study in detail how PLT protein gradients are defined, we first investigated the timescale and levels at which PLT expression is controlled by auxin. Prolonged auxin (indole acetic acid (IAA)) treatment rapidly induced the auxin response marker DR5:GFP<sup>17</sup>, especially when combined with the auxin transport inhibitor, 1-*N*-naphthylphthalamic acid (NPA), but the expression domain of PLTs failed to expand rapidly (Fig. 2a, b and Extended Data Fig. 2a–c). Only after prolonged IAA plus NPA treatment (24–72 h) did expression of PLT–YFPs and the quiescent centre stem cell organizer marker pWOX5:GFP shift shootward, mostly in the meristematic ground tissue (Fig. 2b and Extended Data Fig. 2a–c). This was associated with morphological changes, suggesting that the new PLT expression domain correlated with cell fate changes similar to those described for prolonged NPA treatment<sup>3</sup>. Our experiments thus indicated that PLT induction requires prolonged high auxin levels. To test the implications of these findings, we developed a simulation model of root zonation. The model incorporates a description of root tissue architecture, a generalized PLT–ARF gene regulatory network, root PIN-FORMED (PIN) protein patterns governing auxin transport, and cell growth, division, expansion and differentiation. The resulting model ('initial' model; see Supplementary

Notes, Supplementary Methods and Extended Data Fig. 3) predicts a PLT gradient with shorter range due to its dependence on high auxin levels, in disagreement with experimental observations (Fig. 2c, Supplementary Video 1 and Extended Data Fig. 4). Moreover, *aux1, ein2, gnom* triple mutants, which display a more shallow auxin gradient along the root tip as inferred from direct auxin and auxin response measurements<sup>18</sup>, nevertheless possess a normal range PLT2–YFP gradient (Extended Data Fig. 2d). Together, this demonstrates that the PLT protein gradient is not a direct readout of the auxin gradient.

We investigated how the experimentally observed long PLT protein gradient could arise despite the narrow, non-graded expression domain predicted by our model. One potential explanation emerged when we noticed that PLT2–YFP expression in *pAHP6:XVE>>PLT2-YFP* lines did not only appear in the narrow AHP6 transcription domain (erGFP (where erGFP is a variant of GFP localized to the endoplasmic reticulum) in Fig. 3a), but also in the neighbouring cells (PLT2–YFP in Fig. 3a), suggesting that the protein might influence gradient shape by acting as a mobile plant transcription factor (for a review of this topic, see ref. 19). To ascertain this, we introduced red fluorescent protein (RFP)-tagged PLT2 into a clonal activation system<sup>20</sup> and generated small clones of PLT2–RFP-expressing cells in the meristem. After induction, PLT2–RFP not only resided in clones (marked with green fluorescence) but also in 1–2 cells surrounding the clones (Fig. 3b). When the clones entered the elongation zone, the cells in the clone and the adjacent PLT2–RFP cells remained meristematic and failed to expand ( $n = 7$ ), while cells shootward and rootward from the clone ceased cell division and expanded (Extended Data Fig. 5a–e and Fig. 3b). These data demonstrate that either PLT2 protein or *PLT2* transcript moves to the adjacent cells, yielding translocated functional PLT2–RFP. In addition, the clonal data demonstrate that the inhibition of cell expansion is not the result of a community effect, in which cells in a larger longitudinal region collectively determine whether to expand, but an effect of local PLT levels within the cell file. Fusion of three copies of YFP to PLT2 significantly constrained intercellular movement (Fig. 3a), and when PLT2–3×YFP was expressed under the *PLT2* promoter it complemented the stem cell defect of *plt1,2*, but led to a shorter meristem than when PLT2–YFP was used, indicating that PLT cell-to-cell movement contributes to meristem size (Supplementary Notes and Extended Data Fig. 5f–h).

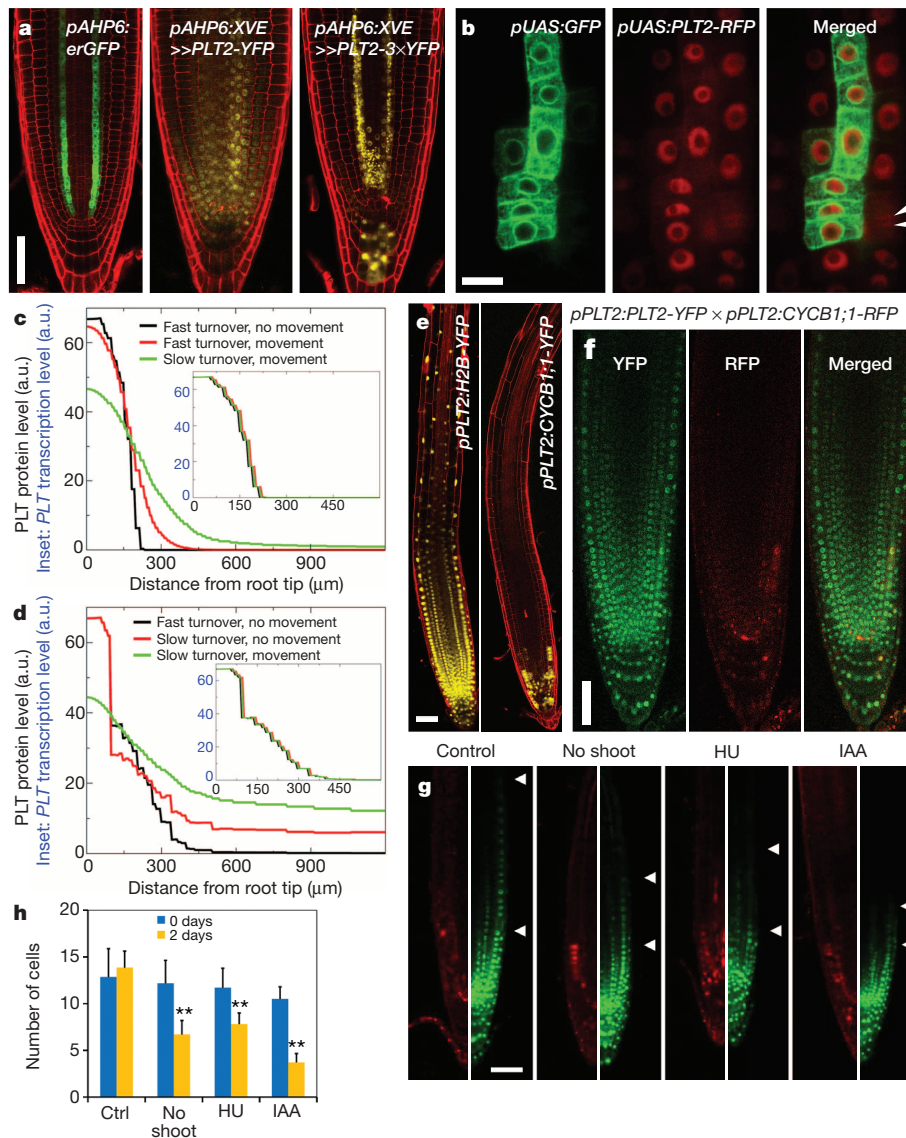
We next performed simulations to analyse how cell-to-cell movement contributed to the PLT gradient. We first simulated PLT movement in the absence of growth and found that, for effective movement, PLT proteins needed to have slow turnover dynamics (Fig. 3c, simulated half-life of ~16 h; see Supplementary Notes and Supplementary Methods). Next, we reinstated root growth. Interestingly, the model predicted that slow PLT turnover in itself substantially contributes to the spread of PLT protein through growth dilution (Fig. 3d).

These new data and previous findings<sup>21</sup> about a regulator of PLT stability highlight a role for protein stability in gradient formation. The previously reported similarity between translational and transcriptional reporter fusion gradients<sup>9</sup> may therefore rather be explained by similar stability of the PLT proteins fused to fluorescence reporters or reporters on their own. To test the influence of protein stability on gradient formation, we used stable and labile proteins fused to the YFP reporter. Histone 2B (H2B), a component of nucleosomes, was used as a stabilizing protein tag, while CYCB1;1, which is degraded from anaphase to S phase<sup>22</sup> in rapidly dividing meristem cells, was employed as a labile protein tag. When driven by the *PLT2* promoter, H2B–YFP displayed fluorescence well into the differentiation zone with a shallow gradient, whereas CYCB1;1–YFP was only present in a punctuate pattern close to the stem cell niche (Fig. 3e). Our data imply that *PLT* genes are transcribed proximal to the stem cell niche, in line with our model predictions, and that retention of PLT proteins in more shootward cells depends critically on their stability. By crossing *pPLT2:CYCB1;1-RFP* with *pPLT2:PLT2-YFP*, we estimated that the *PLT2* transcription domain encompasses approximately one-third of the visible PLT2 protein gradient (Fig. 3f). A subset of the cells in the remaining two-thirds of the PLT2



**Figure 2 | The PLT2 gradient is not a fast readout of the auxin gradient.** **a, b**, Four-day-old seedlings transferred to agar plates containing 20 μM NPA plus 5 μM IAA for the indicated times. Auxin response reporter DR5:erGFP (**a**) rapidly responds to treatment whereas *pPLT2:PLT2-YFP* (**b**) accumulates later (white arrow), associated with repatterning (inset in **b**, magnified image of bracketed region with altered cell division planes at arrowheads). Black arrow indicates fluorescent region after NPA plus IAA treatment (**a**), but not after IAA treatment (Extended Data Fig. 2b). Observed phenotypes/number of roots analysed is indicated in the right bottom corners. **c**, Failure of PLT gradient formation in the initial model. Snapshots of auxin (Aux), *PLT* transcription (*PLT*) and PLT protein (PLT) profiles under steady-state root growth dynamics are displayed. a.u., arbitrary units. Scale bars, 50 μm.





**Figure 3 | Gradient formation by PLT2 cell-to-cell movement and mitotic segregation.** **a**, *AHP6* promoter–*erGFP* fusion is consistently active in two vascular strands (and occasionally in columella). *PLT2-YFP* driven under inducible *AHP6* promoter spreads from its transcription domain, especially in the stem cell region, whereas the movement-deficient version, *PLT2-3×YFP*, is predominantly confined to the *AHP6* transcription domain, although weak signal resides in the stem cell region and between the two vascular strands. **b**, *PLT2-RFP* moves from clone (marked with *GFP*) to neighbouring cells. Arrowheads indicate recently divided nuclei. **c**, **d**, Influence of *PLT* cell-to-cell movement and turnover dynamics on vascular *PLT* protein profiles (main graph) and transcription profiles (inset) in the initial model in the absence

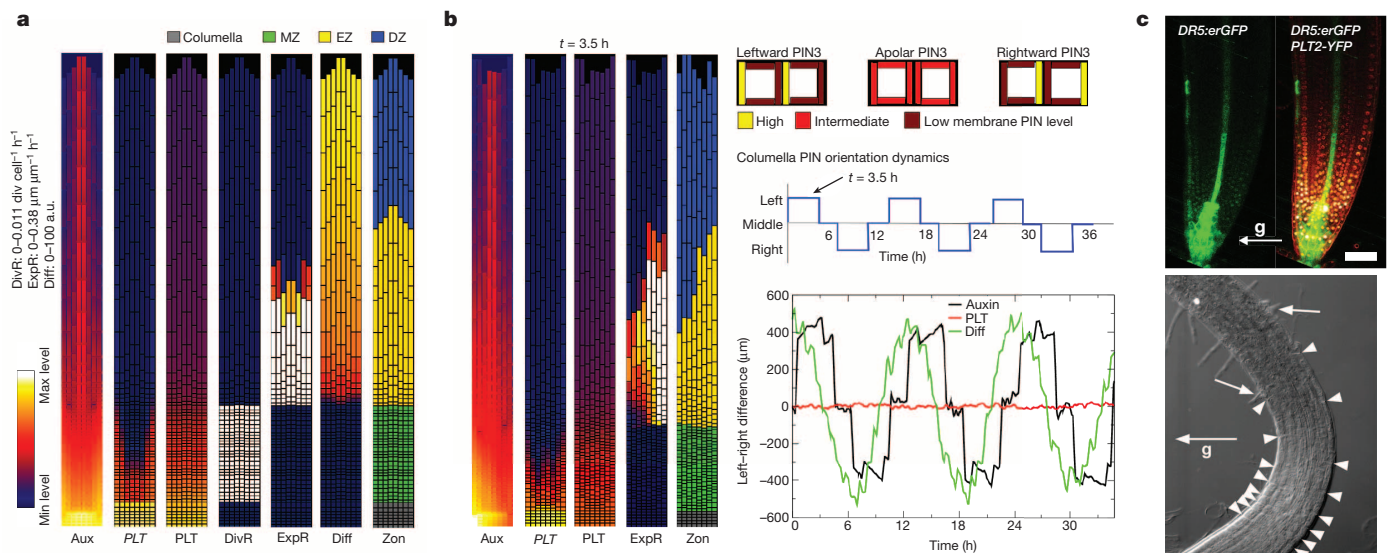
(**c**) or presence (**d**) of growth. **e**, The stability of reporter protein fusion determines the expression pattern driven by the *PLT2* promoter. Stable *H2B-YFP* extends into the differentiation zone, whereas labile *CYCB1;1-YFP* is confined to the meristem. **f**, *PLT2* transcription is in proximal meristem only (red), whereas *PLT2* protein (green) resides in the whole meristem. **g**, **h**, Division inhibition by shoot removal ('No shoot'), HU and IAA treatments shorten *PLT2-3×YFP* gradient. The number of visible YFP-only cells between two arrowheads in **g** are presented in **h**. Ctrl, control. *n* = 7 roots for all treatments, except 6 roots for IAA. Error bars show standard deviation (s.d.). \*\**P* ≤ 0.01, two-way ANOVA with Bonferroni correction. Scale bars, 50 μm, except in **b**, 10 μm.

gradient underwent mitosis, indicating that cells containing *PLT2* protein but not transcribing *PLT2* themselves are still capable of dividing (Supplementary Notes and Extended Data Fig. 5i). Our modelling predicted that besides cell-to-cell movement, growth dilution of *PLT2* by cell division also has a role in the formation of the gradient. To test this, we blocked cell division using IAA (Supplementary Notes), hydroxyurea (HU)<sup>23</sup> or by removing the shoot<sup>4</sup> in *pPLT2:CYCB1;1-RFP* × *pPLT2:PLT2-3×YFP* double reporter lines. We discovered that while the *PLT2* transcription domain remained essentially unaltered, the domain only containing *PLT2-3×YFP* protein was reduced (Fig. 3g, h), confirming a role for growth dilution in gradient formation.

In our simulation model, the incorporation of both root growth and *PLT* intercellular movement with realistic parameter values was necessary to generate a smooth *PLT* gradient capable of dosage-dependent

control of root zonation similar to our experimental observations (Fig. 4a and Supplementary Video 2). Interestingly, a similar gradient-forming mechanism functions in vertebrate axial patterning. There, polarized growth creates a gradient of stable *FGF8* messenger RNA, with diffusion-mediated spread of the *FGF8* protein smoothing and further extending the protein gradient<sup>24</sup>, and *FGF8* itself controlling the growth process<sup>25</sup>. This regulatory architecture, in which growth controls gradient formation and gradient formation controls growth, has been suggested as a robust means to coordinate growth and patterning in polar growing tissues<sup>26</sup>, possibly explaining why it evolved independently in both plants and animals.

Previous studies have suggested roles for auxin in cell division, expansion and differentiation. However, the role of auxin in these processes could only be indirect, through regulation of *PLT* levels. To test



**Figure 4 | Root zonation under normal growth and gravitropism.**

**a**, Zonation dynamics in the PLT-spread model under normal growth conditions. Snapshots of the auxin distribution (Aux), PLT transcription (*PLT*), PLT protein (PLT), division rate (DivR, measured as number of divisions per cell per hour), cell expansion rate (ExpR, measured as growth ( $\mu\text{m}$ ) per unit tissue ( $\mu\text{m}$ ) per hour), differentiation level (Diff) and zonation dynamics (Zon) profiles. **b**, Root zonation dynamics in the gravitropism model under dynamic gravitropism. Left, snapshots of auxin, *PLT* transcription, PLT protein, expansion rate and zonation for leftward oriented gravity vector. (For downward and rightward oriented gravity vector see Extended Data Fig. 9b.)

this hypothesis, we next investigated whether there is also a direct role for auxin in controlling root zonation dynamics. To focus on direct effects of auxin, we considered short timescales insufficient to lead to changes in PLT expression. Auxin addition, application of the auxin antagonist auxinole<sup>27</sup>, and inhibition of auxin signalling by inducing the stable ARF-signalling repressor *axr3-1* (ref. 28) experiments all confirmed that auxin rapidly regulates all zonation processes. Cell division and expansion rates depended on optimum auxin levels, with different thresholds, whereas differentiation required a minimum level of auxin (see Supplementary Notes, Supplementary Videos 3, 4 and Extended Data Figs 6 and 7). Our computational model could readily be extended with these auxin-dependent rates ('auxin model'), reproducing both normal zonation and the experiments described earlier (see Supplementary Notes, Supplementary Methods and Extended Data Fig. 8).

Thus, our study uncovered a regulatory architecture in which auxin: (1) rapidly influences rates of developmental processes within zones without directly affecting PLT levels (minutes to hours timescale); and (2) influences the size and location of differentiation zones slowly through regulating *PLT* transcription (timescale of days). A subtle coupling between these processes occurs because auxin influences PLT growth dilution through division and expansion rates (timescale of hours) and hence the location of the division and expansion thresholds (Extended Data Figs 2c, 6a, 8c, Supplementary Notes and Supplementary Methods). The coexistence of slow, PLT-mediated and rapid, direct auxin effects on zonation made us wonder why such an elaborate control system has evolved. To investigate this, we analysed gravitropism, an auxin-mediated process operating at a faster timescale than the generation of the PLT gradient. Gravity stimuli drive PIN protein reorientation-mediated asymmetric auxin accumulation on the lower side of the root within 5 min (refs 13, 14), causing inhibition of cell expansion, and bending of the root towards the new gravity vector within 6 h (refs 13, 14) (Fig. 4c). When PIN protein reorientation caused by alternating gravitropic stimuli was simulated in our model ('gravitropism model', Fig. 4b), elevated auxin levels alternated from left to right in the root and induced the differential expansion that drives root bending, while PLT levels stayed constant (Fig. 4b, Supplementary Video 5 and Extended Data Fig. 9a–d).

Right, dynamics of left–right differences in auxin, differentiation level and PLT protein distribution. Depicted are the used columella PIN orientation patterns, the applied 12 h cycle of PIN orientation changes, and the resulting auxin, differentiation level and PLT left–right distribution differences (see Supplementary Methods). *t* indicates the time point at which the snapshot was taken. **c**, DR5 and PLT expression in the same root (top) after gravitropic stimulation resulting in left–right difference in appearance of the first root hair (bottom). Arrows with 'g', gravity vector; white arrowheads, individual cells in the elongation zone; white arrows, youngest root hairs. Scale bar, 50  $\mu\text{m}$ .

The predicted constant PLT levels were confirmed experimentally (Fig. 4c). Thus, this regulatory design allows for a partial separation of timescales that enables rapid auxin-mediated tropic responses, essential for sessile plants to respond to environmental challenges, while maintaining stable PLT-mediated developmental zonation (Extended Data Fig. 10a–c and Supplementary Discussion). If, in contrast, as was previously thought, PLT expression were a relatively direct and proportionate readout of auxin levels, both auxin and PLT patterns would fluctuate under tropisms, resulting in variable zonation patterns and loss of co-ordinated differentiation (Extended Data Fig. 10d, e and Supplementary Discussion).

We uncover the auxin–PLT network as a core module on which other factors, such as other phytohormones (for a review, see ref. 29), can act to regulate growth. Our study prompts two directions for future exploration. First, recently uncovered positive feedbacks from PLT back to auxin biosynthesis and transport<sup>9,10,30</sup> do not notably affect the behaviour of our model (Extended Data Fig. 9e–g). We speculate that these feedbacks may have a role only during the generation of new primordia, when robust, localized auxin and PLT maxima need to be established. Second, the dominant role of PLT gradients in controlling zonation dynamics challenges the role of an auxin gradient as a dose-dependent instructive signal. Indeed, recent studies suggest that the auxin profile may not be a simple gradient<sup>6,11</sup>. While our results support a role for auxin levels in zonation, they leave undecided whether a specific gradient-shaped auxin distribution is required.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 6 February; accepted 8 July 2014.

Published online 24 August 2014.

- Dolan, L. *et al.* Cellular organisation of the *Arabidopsis thaliana* root. *Development* **119**, 71–84 (1993).
- Beemster, G. T. & Baskin, T. I. Stunted plant 1 mediates effects of cytokinin, but not of auxin, on cell division and expansion in the root of *Arabidopsis*. *Plant Physiol.* **124**, 1718–1727 (2000).

3. Sabatini, S. *et al.* An auxin-dependent distal organizer of pattern and polarity in the *Arabidopsis* root. *Cell* **99**, 463–472 (1999).
4. Grieneisen, V. A., Xu, J., Maree, A. F., Hogeweg, P. & Scheres, B. Auxin transport is sufficient to generate a maximum and gradient guiding root growth. *Nature* **449**, 1008–1013 (2007).
5. Petersson, S. V. *et al.* An auxin gradient and maximum in the *Arabidopsis* root apex shown by high-resolution cell-specific analysis of IAA distribution and synthesis. *Plant Cell* **21**, 1659–1668 (2009).
6. Brunoud, G. *et al.* A novel sensor to map auxin response and distribution at high spatio-temporal resolution. *Nature* **482**, 103–106 (2012).
7. Ishida, T. *et al.* Auxin modulates the transition from the mitotic cycle to the endocycle in *Arabidopsis*. *Development* **137**, 63–71 (2010).
8. Aida, M. *et al.* The PLETHORA genes mediate patterning of the *Arabidopsis* root stem cell niche. *Cell* **119**, 109–120 (2004).
9. Galinha, C. *et al.* PLETHORA proteins as dose-dependent master regulators of *Arabidopsis* root development. *Nature* **449**, 1053–1057 (2007).
10. Bliou, I. *et al.* The PIN auxin efflux facilitator network controls growth and patterning in *Arabidopsis* roots. *Nature* **433**, 39–44 (2005).
11. Band, L. R. *et al.* Systems analysis of auxin transport in the *Arabidopsis* root apex. *Plant Cell* **26**, 862–875 (2014).
12. Hofhuis, H. *et al.* Phyllotaxis and rhizotaxis in *Arabidopsis* are modified by three PLETHORA transcription factors. *Curr. Biol.* **23**, 956–962 (2013).
13. Friml, J., Wisniewska, J., Benkova, E., Mendgen, K. & Palme, K. Lateral relocation of auxin efflux regulator PIN3 mediates tropism in *Arabidopsis*. *Nature* **415**, 806–809 (2002).
14. Band, L. R. *et al.* Root gravitropism is regulated by a transient lateral auxin gradient controlled by a tipping-point mechanism. *Proc. Natl Acad. Sci. USA* **109**, 4668–4673 (2012).
15. Mähönen, A. P. *et al.* Cytokinin signaling and its inhibitor AHP6 regulate cell fate during vascular development. *Science* **311**, 94–98 (2006).
16. Lee, M. M. & Schiefelbein, J. WEREWOLF, a MYB-related protein in *Arabidopsis*, is a position-dependent regulator of epidermal cell patterning. *Cell* **99**, 473–483 (1999).
17. Friml, J. *et al.* Efflux-dependent auxin gradients establish the apical–basal axis of *Arabidopsis*. *Nature* **426**, 147–153 (2003).
18. Fischer, U. *et al.* Vectorial information for *Arabidopsis* planar polarity is mediated by combined AUX1, EIN2, and GNOM activity. *Curr. Biol.* **16**, 2143–2149 (2006).
19. Wu, S. & Gallagher, K. L. Transcription factors on the move. *Curr. Opin. Plant Biol.* **15**, 645–651 (2012).
20. Heidstra, R., Welch, D. & Scheres, B. Mosaic analyses using marked activation and deletion clones dissect *Arabidopsis* SCARECROW action in asymmetric cell division. *Genes Dev.* **18**, 1964–1969 (2004).
21. Matsuzaki, Y., Ogawa-Ohnishi, M., Mori, A. & Matsubayashi, Y. Secreted peptide signals required for maintenance of root stem cell niche in *Arabidopsis*. *Science* **329**, 1065–1067 (2010).
22. Colón-Carmona, A., You, R., Haimovitch-Gal, T. & Doerner, P. Technical advance: spatio-temporal analysis of mitotic activity with a labile cyclin–GUS fusion protein. *Plant J.* **20**, 503–508 (1999).
23. Culligan, K., Tissier, A. & Britt, A. ATR regulates a G2-phase cell-cycle checkpoint in *Arabidopsis thaliana*. *Plant Cell* **16**, 1091–1104 (2004).
24. Yu, S. R. *et al.* Fgf8 morphogen gradient forms by a source-sink mechanism with freely diffusing molecules. *Nature* **461**, 533–536 (2009).
25. Wilson, V., Olivera-Martinez, I. & Storey, K. G. Stem cells, signals and vertebrate body axis extension. *Development* **136**, 1591–1604 (2009).
26. Ibañez, M., Kawakami, Y., Rasskin-Gutman, D. & Izpisua Belmonte, J. C. Cell lineage transport: a mechanism for molecular gradient formation. *Mol. Syst. Biol.* **2**, 57 (2006).
27. Hayashi, K. *et al.* Rational design of an auxin antagonist of the SCF<sup>TIR1</sup> auxin receptor complex. *ACS Chem. Biol.* **7**, 590–598 (2012).
28. Rouse, D., Mackay, P., Stirnberg, P., Estelle, M. & Leyser, O. Changes in auxin response from mutations in an *AUX/IAA* gene. *Science* **279**, 1371–1373 (1998).
29. Vanstraelen, M. & Benkova, E. Hormonal interactions in the regulation of plant development. *Annu. Rev. Cell Dev. Biol.* **28**, 463–487 (2012).
30. Pinon, V., Prasad, K., Grigg, S. P., Sanchez-Perez, G. F. & Scheres, B. Local auxin biosynthesis regulation by PLETHORA transcription factors controls phyllotaxis in *Arabidopsis*. *Proc. Natl Acad. Sci. USA* **110**, 1107–1112 (2013).

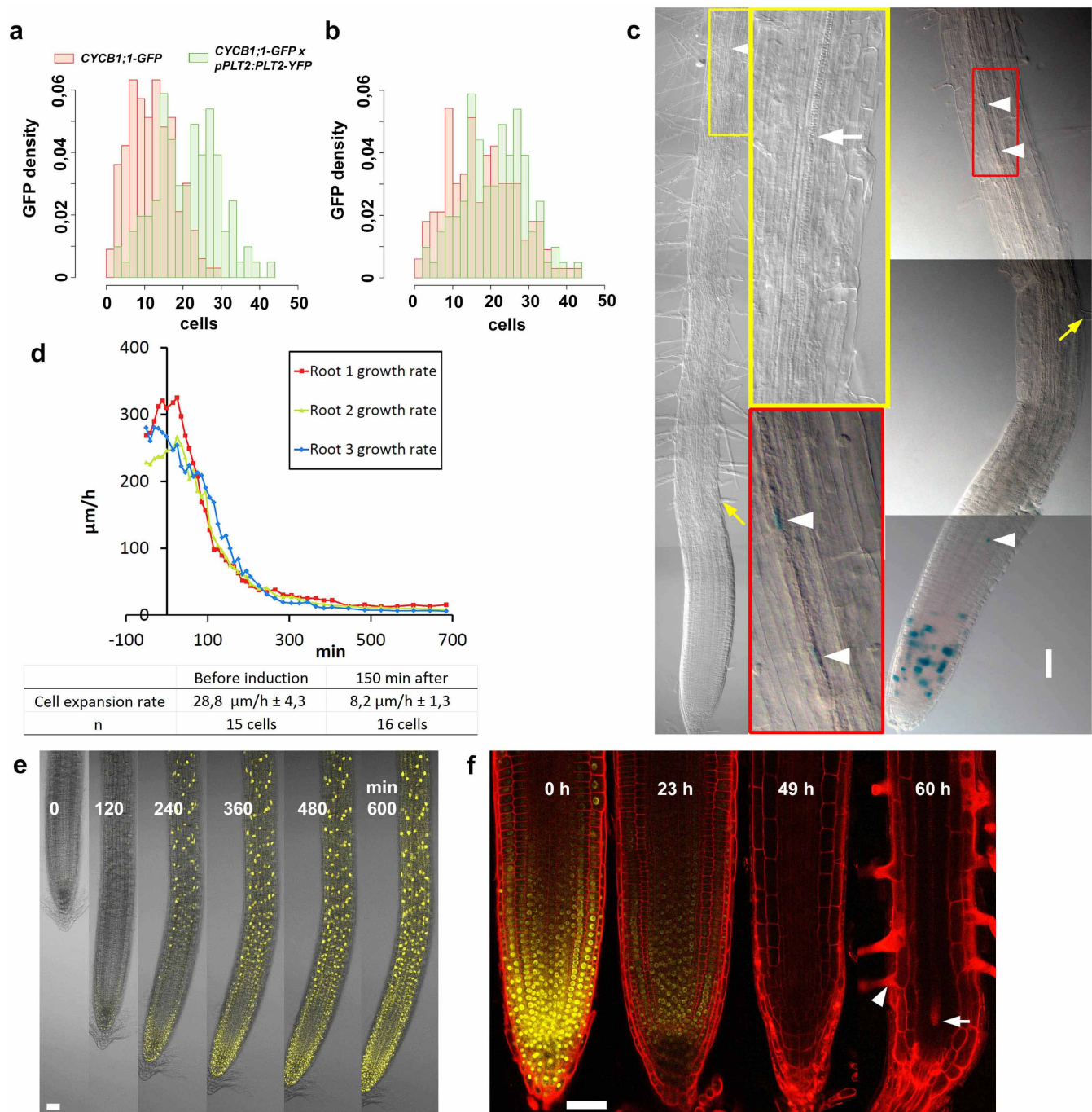
**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** We thank M. Grebe, P. Benfey and K.-i. Hayashi for materials; the Light Microscopy Unit (Institute of Biotechnology), S. El-Showk, A.-M. Bågman, J. van Amerongen and F. Kindt for technical advice or assistance. This work is supported by a Human Frontier Science Program fellowship (A.P.M.), European Research Council Advanced Investigator Fellowship SysArc (B.S.), SPINOZA award (B.S., K.t.T., K.P.), ALW-NWO European Research Area Network Plant Genomics (ERAPG) grant 855.50.017 (S.D.-T.), the Academy of Finland (A.P.M., R.S., O.S., J.S.), Biocentrum Helsinki and University of Helsinki (A.P.M., R.S., O.S.), Integrative Life Science Doctoral Program (R.S.), Marie Curie Intra-European Fellowship (IEF-2008-237643) (S.D.-T.), the Netherlands Organisation for Scientific Research (NWO)-Horizon grant (R.H.), NWO-ALW grant (G.W.), and EMBO Long-term fellowship (A.P.M., K.P.).

**Author Contributions** A.P.M. and B.S. designed the experiments. A.P.M., R.S., O.S. and S.D.-T. carried out the experiments. K.t.T. designed and performed computational simulations. J.S. performed statistical analyses. G.W., K.P. and R.H. provided material for the study. A.P.M., K.t.T. and B.S. wrote the manuscript.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.P.M. (AriPekka.Mahonen@helsinki.fi) or B.S. (ben.scheres@wur.nl).

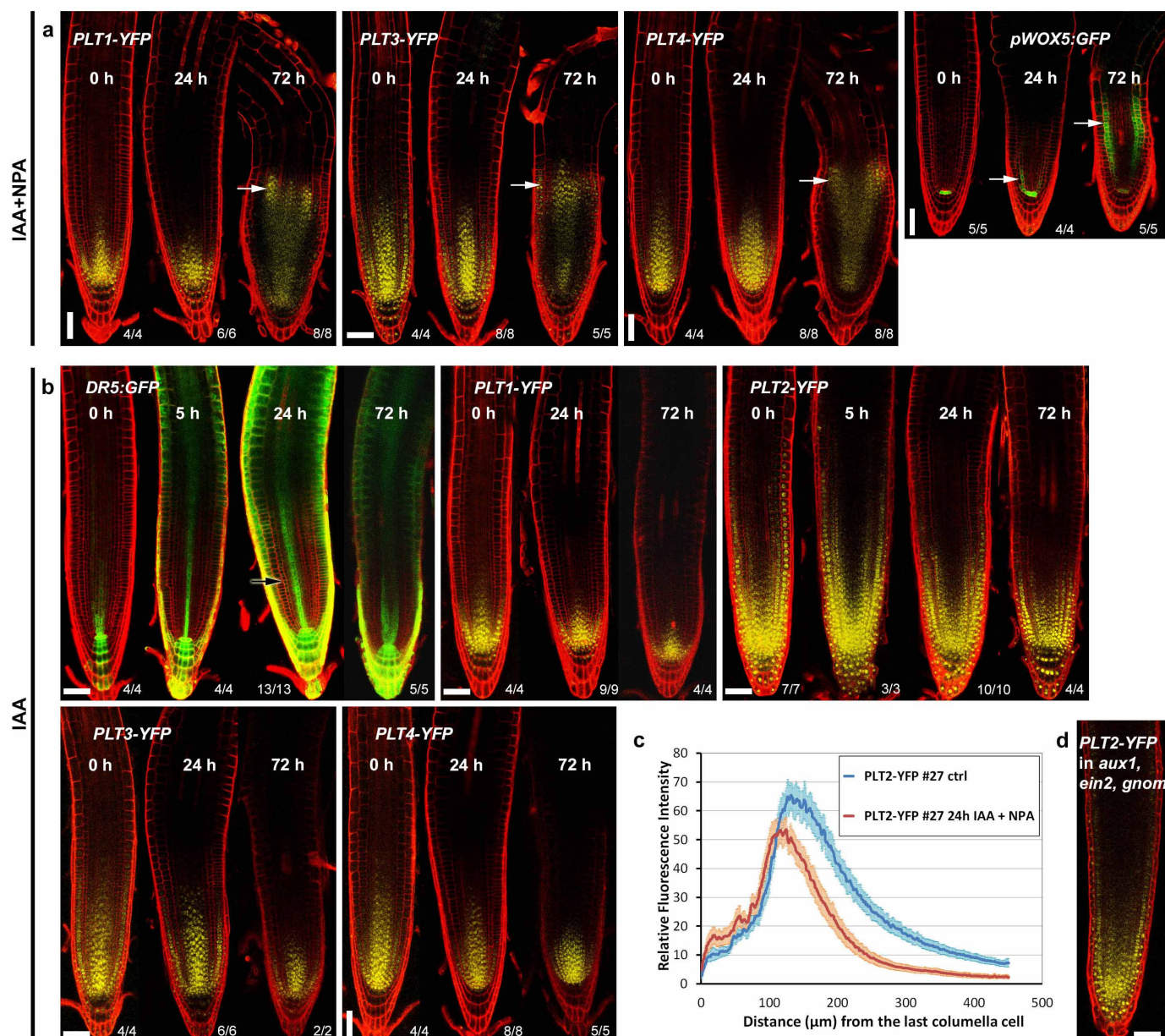




### Extended Data Figure 1 | PLTs are dose-dependent drivers of zonation.

**a, b**, The domain of frequent cell division, monitored by cell cycle marker *CYCB1;1-GFP* in Fig. 1b, c, shifts shootward with increased *PLT2* dosage (that is, homozygote *pPLT2:PLT2-YFP* in Col background). Histogram in **a** shows the distribution of the *CYCB1;1-GFP*-positive cells along the meristem at a given distance from the quiescent centre. *x* axis indicates the distance from the quiescent centre as number of cortical cells, and *y* axis label 'GFP density' refers to the proportion of *CYCB1;1-GFP*-containing cells at the given distance from the quiescent centre. Shootward shift of the distance of the cell division events in the presence of increased *PLT2* (green histogram) dosage compared to wild-type (red histogram) is significant (*t*-test for mean  $P \ll 0.001$ , Wilcoxon test for median  $P \ll 0.001$ , Kolmogorov–Smirnov for difference of distributions  $P \ll 0.001$ ). **b**, Histogram presenting rescaled data to show that the distribution of the high cell division domain shifted shootward when *PLT2* dosage was increased. A null hypothesis was that shootward shift is due to higher dispersion of the distribution observed under increased *PLT2* dosage. To test this hypothesis, the control *CYCB1;1-GFP* data were rescaled to match the maximum values of *PLT2* data. The null hypothesis was rejected (*t*-test

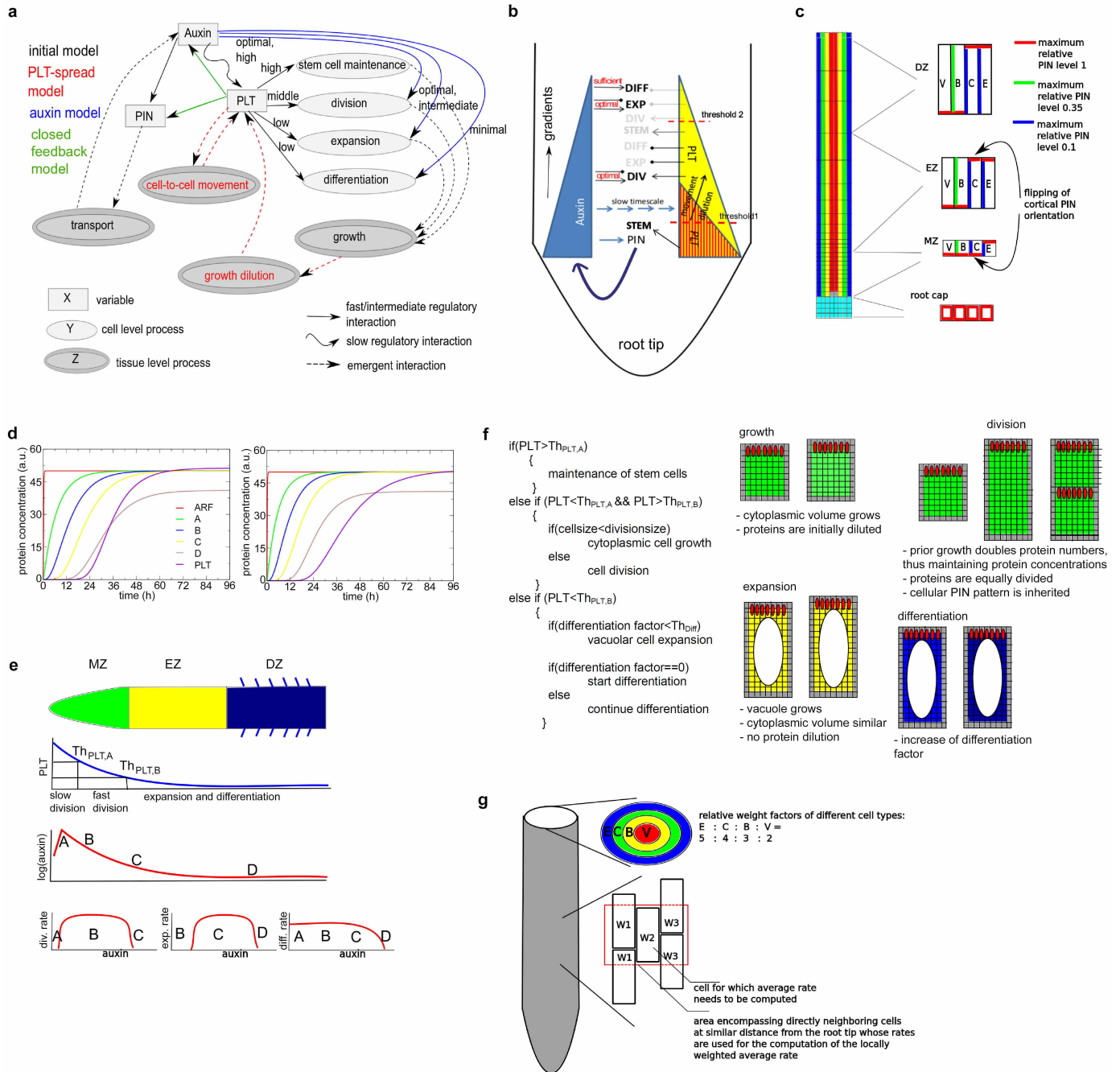
$P = 0.001$ , Wilcoxon test  $P = 0.0012$ , Kolmogorov–Smirnov  $P = 0.0026$ ), indicating that the shootward shift of the high division domain in the presence of increased *PLT2* dosage is significant, and not due to dispersion. The bin width in histograms is two cells (that is, 1st bar, 1 and 2 cells; 2nd, 3 and 4, and so on). **c**, Induction of *pAHP6:XVE*  $\gg$  *PLT2-YFP* inhibits xylem differentiation (left) (white arrow indicates the first protoxylem element) and triggers ectopic cell divisions illustrated by *CYCB1;1-GUS* activity (right) (arrowheads), whereas root hairs develop normally (yellow arrows). Insets show magnifications from the designated areas. **d**, *PLT2-YFP* induction ( $t = 0$ ) rapidly inhibits root growth (top) and cell expansion (bottom) in three roots. Expansion rates as shown as averages  $\pm$  s.d. **e**, Inhibition of growth coincides with appearance of *PLT2-YFP* signal after induction. **f**, Induction of *PLT2* RNAi (in *plt1,3,4; pPLT2:PLT2-YFP*) abolishes the *PLT2-YFP* signal by 49 h and consequently promotes expansion and differentiation of the meristem cells, as indicated by the appearance of expanded cells as well as protoxylem (arrow) and root hairs (arrowhead) in the meristem. Images from the same root using identical confocal microscopy settings for the yellow channel. Scale bars, 50  $\mu\text{m}$ .



**Extended Data Figure 2 | PLT expression patterns respond only to long-term auxin accumulation in the meristem.** **a**, PLT expression shifts shootward only when prolonged auxin application is accompanied with polar auxin transport inhibitor (NPA) treatment. Four-day-old seedlings were transferred to an agar plate containing 20  $\mu\text{M}$  NPA plus 5  $\mu\text{M}$  IAA for the time periods indicated in the images. The expression of *pPLT1:PLT1-YFP*, *pPLT3:PLT3-YFP* and *pPLT4:PLT4-YFP* spreads shootward (white arrows) by 72 h of NPA plus IAA treatment. **b**, PLT expression patterns are insensitive for auxin-only treatments. Four-day-old seedlings were transferred to an agar plate containing 5  $\mu\text{M}$  IAA for the time periods indicated in the images. The auxin response reporter *DR5:erGFP* rapidly responded to the treatment whereas the expression domains of *pPLT1:PLT1-YFP*, *pPLT2:PLT2-YFP*, *pPLT3:PLT3-YFP* and *pPLT4:PLT4-YFP* failed to expand. Black arrow

indicates the region in meristem that is absent of *DR5:erGFP* fluorescence after IAA treatment but is filled with fluorescence after NPA plus IAA treatment (Fig. 2a). Observed phenotypes/number of roots analysed are indicated in the right bottom corners. **c**, Twenty-four hours of NPA plus IAA treatment fails to expand the *PLT2-YFP* gradient shootward. In fact, the treatment leads to transient shortening of the *PLT2-YFP* gradient, probably due to inhibition of growth dilution of *PLT2-YFP* in the meristematic cells (see Fig. 3g, h).  $P \ll 0.001$ , Kolmogorov-Smirnov test; error bars indicate 95% confidence intervals.  $n = 20$  (dimethylsulphoxide (DMSO)) and 23 (NPA plus IAA). **d**, Graded *pPLT2:PLT2-YFP* expression despite shallow auxin gradient in *aux1, ein2, gnom*; representative image from three independent lines. Scale bars, 50  $\mu\text{m}$ .

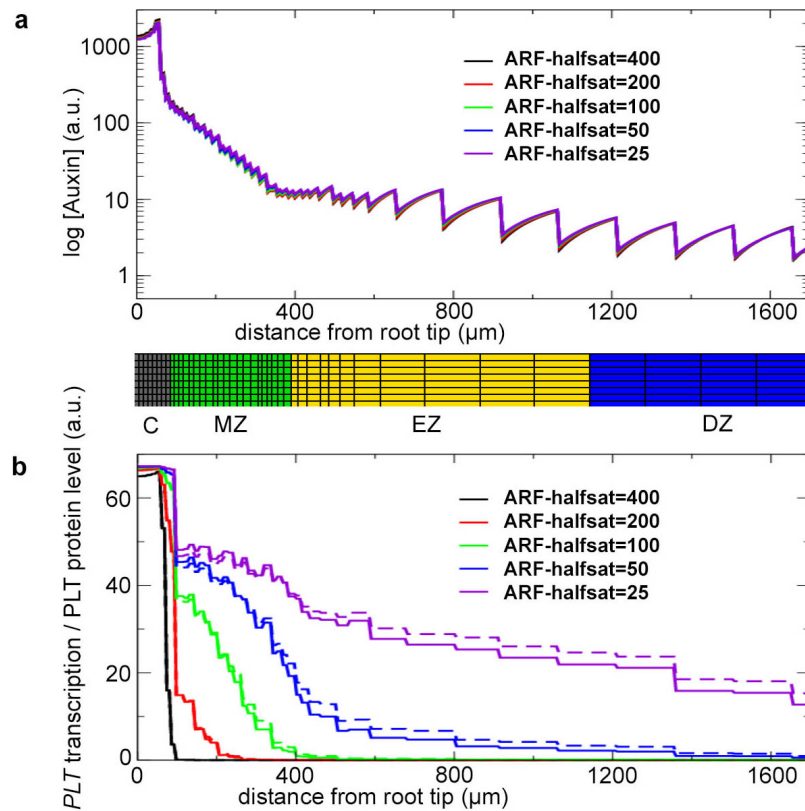






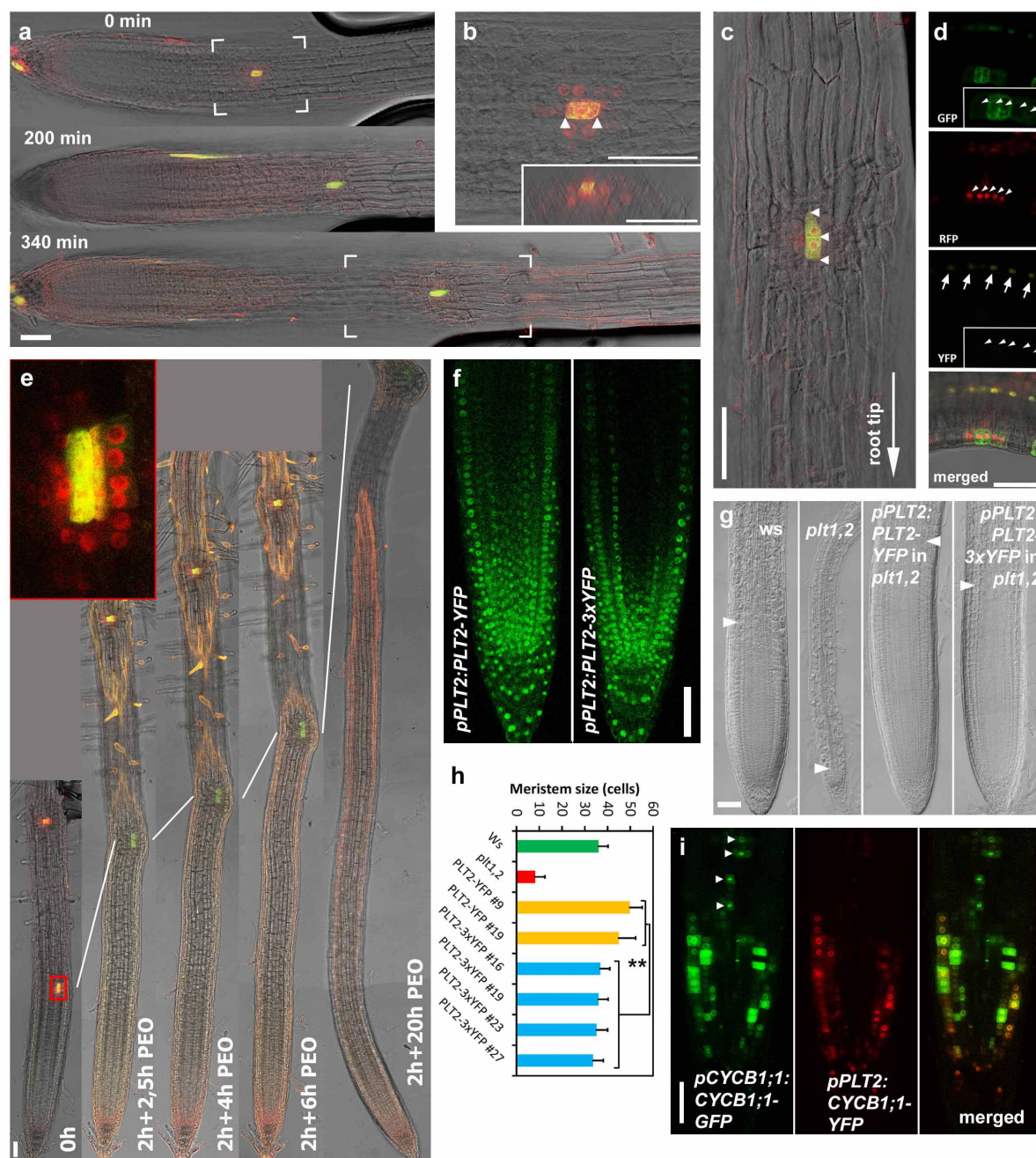
**Extended Data Figure 3 | Composition of the zonation model.** **a**, Overview of the initial, PLT-spread, auxin and closed feedback models. Relationships between model variables (grid-based auxin levels, cell-based PIN and PLT levels), growth processes (stem cell maintenance, division, expansion and differentiation), PIN-mediated auxin transport, cell-to-cell PLT movement and their emerging consequences for auxin and PLT levels are shown, with different colours indicating from which model onward they are incorporated (models are in order of increasing complexity). Variables are indicated in rectangles, processes are indicated in ovals. A distinction is made between cell-level and tissue-level processes. Direct regulatory interactions are indicated with arrows, and a distinction is made between fast to intermediate speed versus slow interactions, and between the levels of the 'input' variable needed for a particular process to occur. Emergent feedbacks, with processes influencing variable levels other than through a direct regulatory effect (for example, growth spreading PLT and hence influencing its levels) are indicated as dashed arrows. **b**, Spatially explicit overview of the interplay between the variables and processes and how these generate the auxin and PLT gradients and control root zonation dynamics in the auxin model. Stem cells and slow division (STEM), fast division (DIV), expansion (EXP) and differentiation (DIF) processes are indicated in black if the local auxin and PLT levels permit these processes, and in grey if the local auxin and PLT levels prevent these processes from occurring. Similarly, black arrows indicate that local auxin or PLT levels cause promotion or repression of a process, grey arrows indicate that auxin or PLT in principle has a particular effect on a process but that local levels do not allow for this effect to occur. Blue arrows indicate transcriptional effects. The broad dark blue arrow indicates PIN-mediated auxin transport. **c**, Layout of the root tissue and PIN distribution pattern used in the simulations. The simulated tissue contains columella cells (cyan), quiescent centre cells (grey), epidermal cells (blue), cortical cells (green), border cells (yellow) and

vasculature cells (red). For the PIN distribution pattern, four levels are distinguished: no PINs, a maximum relative level of 0.1, 0.35, and 1. These maximum relative levels are multiplied with the cellular PIN protein level to determine a membrane segment's actual PIN level. The rootward-shootward flip of PIN polarity in cortex cells is indicated with arrows. **d**, Time course of free ARF levels, and protein levels of transcription factors A, B, C, D and PLT under a constant auxin application (level 100). Left, fast PLT protein turnover; right, slow PLT protein turnover. **e**, Schematic overview of root zonation and its dependence on PLT and auxin gradients. Top panel, location of the meristem (MZ), expansion (EZ) and differentiation (DZ) zones. Second panel, PLT concentration profile along the length of the root. PLT levels dictate the location of stem cell maintenance and slow division, fast division and expansion and differentiation domains, with the first threshold demarcating the boundary between stem cell and fast division domains and the second threshold demarcating the boundary between fast division and expansion and differentiation domains. Third panel, auxin profile along the root. Fourth panel, dependence of division, expansion and differentiation rates on auxin levels. A, B, C and D correspond to different auxin levels; for comparison purposes it is shown where these levels occur both in the auxin gradient profile (third panel) and where these levels occur in the division, expansion and differentiation auxin rate dependency functions (fourth panel). **f**, Pseudocode of the algorithm used in the program to determine whether cells will grow, divide, expand or differentiate. In addition, a cartoon version of the consequences of growth, division, expansion and differentiation processes on gene expression levels, differentiation levels, cell size and PIN pattern is shown. **g**, Schematic depiction of the neighbouring cells and weighting factors used to calculate weighted, locally averaged growth and expansion rates. B, boundary cells; C, cortex cells; E, epidermal cells; V, vascular cells.



**Extended Data Figure 4 | The requirement of high auxin levels produces narrow, non-graded PLT profiles.** **a**, Vascular auxin concentration profiles under root zonation dynamics in the initial model for different half-saturation values for free ARF. Note: the curves are practically superimposed. A snapshot

image of the zonation below the graph illustrates the location of the root zones. Columella (C), meristem (MZ), expansion (EZ) and differentiation (DZ) zones are shown. **b**, Vascular PLT transcription (continuous lines) and protein profiles (dashed lines) for the same parameter settings as in **a**.

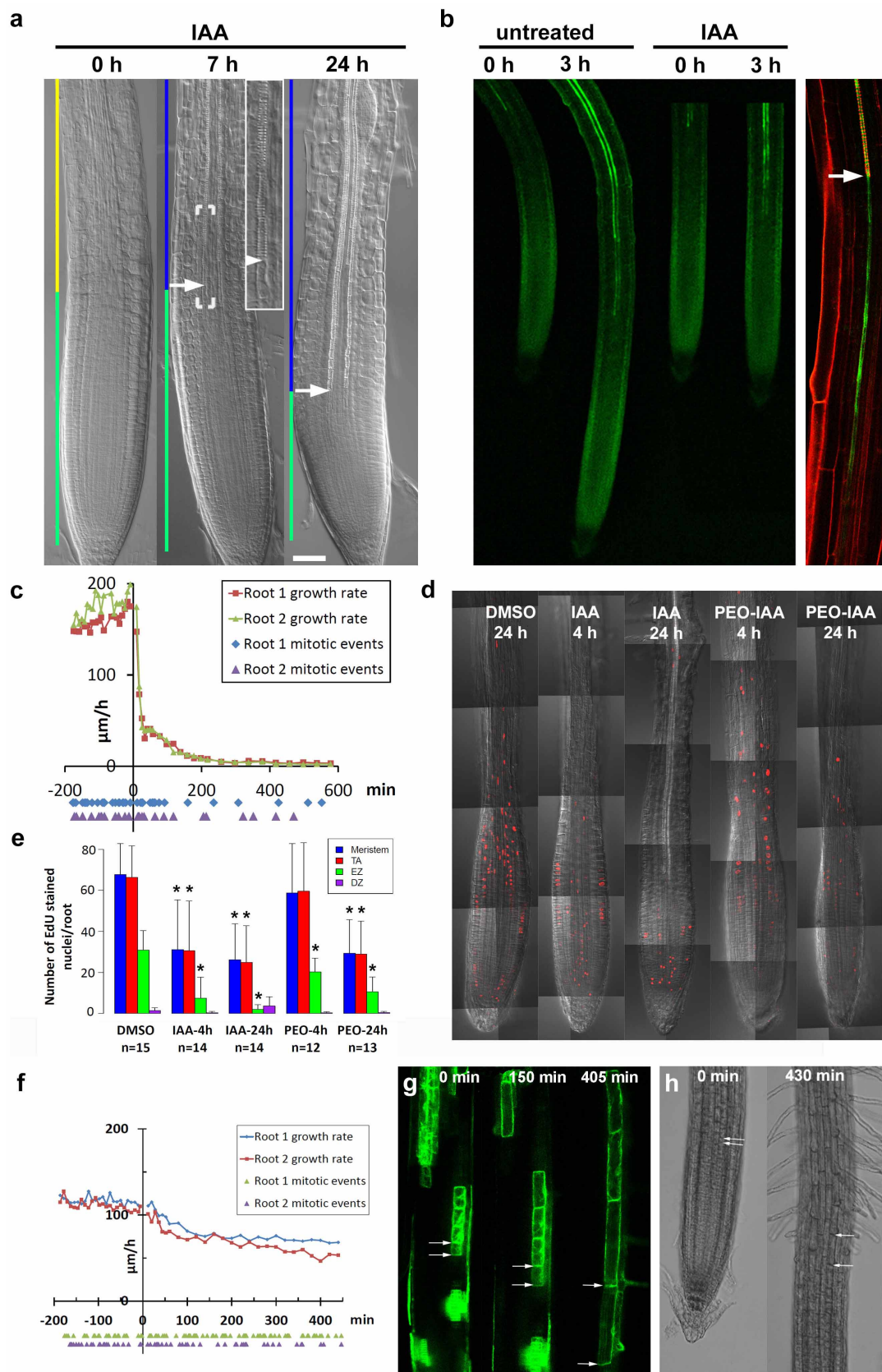


**Extended Data Figure 5 | PLT2 protein persistence and mobility maintain meristematic characteristics without elevating auxin response.** **a**, A single cell clone expressing PLT2-RFP exiting the meristem (0 min) and travelling through the expansion zone towards the differentiation zone (200 min and 340 min). **b**, Magnification of the marked area in **a** (0 min), demonstrating that the clone (marked with green fluorescence, appearing as yellow when overlapped with PLT2-RFP red fluorescence) consists originally of a single cell (white arrowheads). Note that PLT2-RFP (red nuclear fluorescence) is present both in the clone and the surrounding cells. Inset shows optical cross-section at the position of the clone. **c**, Magnification of the marked area in **a** (340 min) showing that the clone has divided once while being in the expansion zone (arrowheads mark two clonal cells), and that PLT2-RFP-expressing cells do not expand, whereas cells produced before and after generation of the clone have expanded. **d**, Auxin response sensor DR5:nYFP (yellow nuclear fluorescence)<sup>31</sup> is not elevated in the PLT2-RFP cells (white arrowheads) but shows normal response in vasculature (white arrows). **e**, Anti-auxin,

$\alpha$ -(phenylethyl-2-oxo)-IAA (PEO-IAA) inhibits root hair formation, but fails to promote cell expansion in the PLT2-RFP clones. A PLT2-RFP clone exiting the meristem (0 h) and travelling through the elongation zone towards the differentiation zone. PEO-IAA (30  $\mu$ M) was applied to the medium 2 h after taking the first (0 h) image. Then images were taken 2.5 h, 4 h, 6 h and 20 h after PEO-IAA application. Note: root hair production is inhibited after PEO-IAA application. Inset, magnification of the marked area in the 0 h image, showing the clone (marked with green fluorescence) and that PLT2-RFP (red nuclear fluorescence) is present both in the clone and the surrounding cells. **f**, PLT2-3 $\times$ YFP shows reduced expression in the stele. **g**, **h**, The movement-deficient version, PLT2-3 $\times$ YFP, complements the *plt1,2* mutant, although the meristem is shorter than when PLT2-YFP is used. Seedlings were 7 days old; arrowheads, MZ boundary. Asterisks in **h**, Wilcoxon test ( $P < 0.001$ ); meristem size of PLT2-3 $\times$ YFP lines significantly reduced. Error bars show s.d. **i**, Cells shootward from the stem cell niche are proliferating without PLT2 transcription (arrowheads, cells with GFP but no RFP).

31. Heisler, M. G. *et al.* Patterns of auxin transport and gene expression during primordium development revealed by live imaging of the *Arabidopsis* inflorescence meristem. *Curr. Biol.* **15**, 1899–1911 (2005).

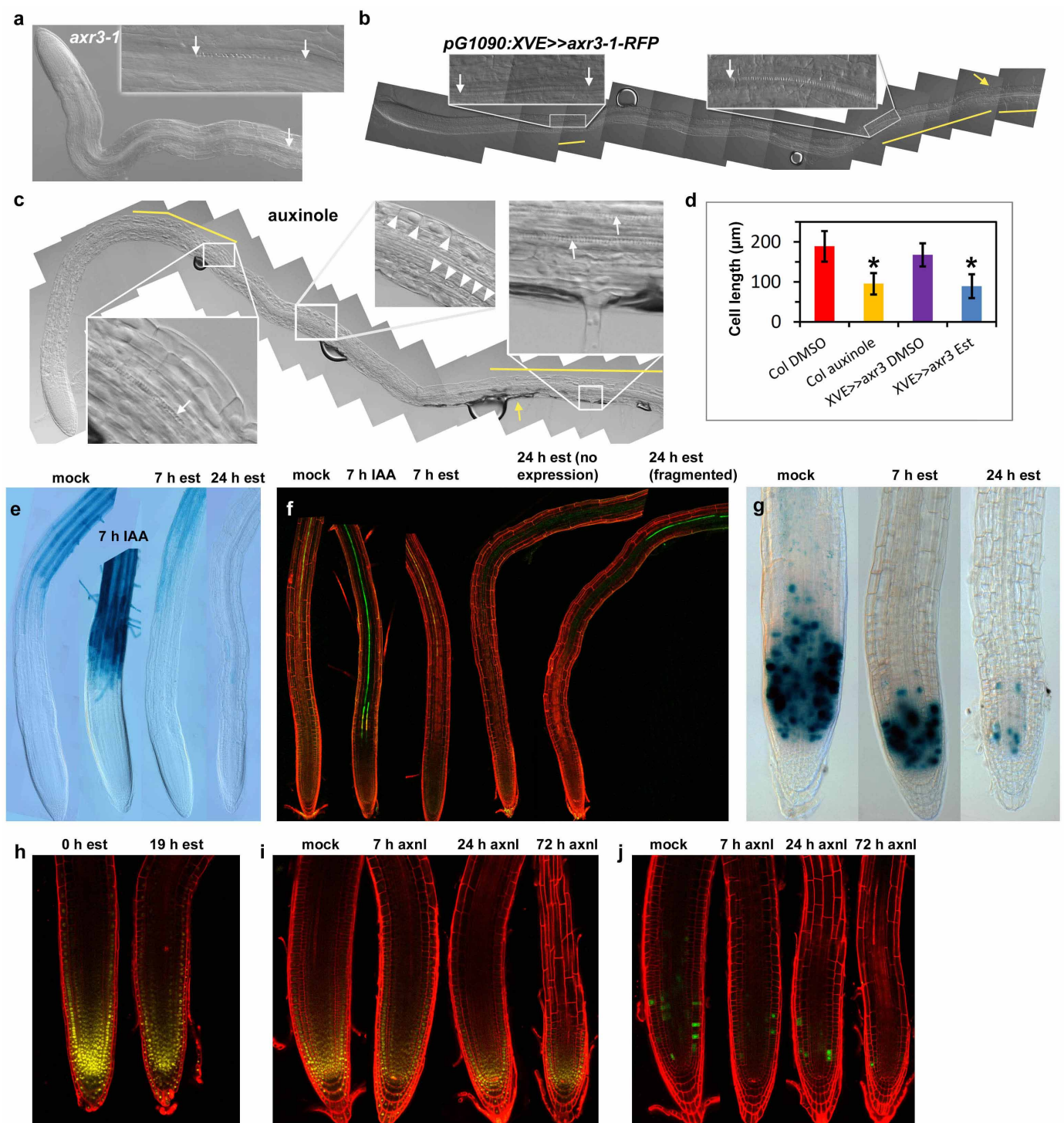




**Extended Data Figure 6 | High auxin levels rapidly inhibit cell division and expansion, but not differentiation.** **a**, First signs of expansion zone differentiation 7 h after 5  $\mu$ M IAA application marked by appearance of protoxylem elements (arrow). By 24 h ubiquitous differentiation of the expansion zone is evident. 0 h image used from Fig. 1a. Scale bar, 50  $\mu$ m. **b**, Auxin application rapidly inhibits growth while xylem differentiation, monitored by green fluorescence of S18 marker<sup>32</sup>, proceeds towards meristem. Snapshots from a video of the same root before and after application. Right panel, S18 signal is tightly associated with protoxylem differentiation (arrow). **c**, Root growth ( $\mu$ m h<sup>-1</sup>) and mitosis (below the  $x$  axis) of two roots over time (min). IAA applied at  $t = 0$ . **d**, **e**, Application of 5  $\mu$ M IAA and inhibition of auxin signalling by 30  $\mu$ M PEO-IAA leads to decreased accumulation of 5-ethynyl-2'-deoxyuridine (EdU) stain (red fluorescence), marking DNA replication. Asterisks in **e**, Mann–Whitney U test  $P < 0.05$ , after Bonferroni correction of multiple comparisons; reduction of number of EdU-stained nuclei compared with DMSO control. Error bars show s.d. **f**, Application of moderate levels of IAA (30 nM) still inhibited cell expansion (Supplementary Notes) but did not inhibit cell division. Root growth ( $\mu$ m h<sup>-1</sup>) and mitotic

events (below the  $x$  axis) of two roots over time (min). IAA (30 nM) was applied at  $t = 0$ . **g**, **h**, To measure the duration of the differentiation process, individual cells were followed as they left the meristem, expanded and entered the differentiation zone. **g**, Tracking of a GFP clone<sup>20</sup> consisting of four cells. Arrows highlight a cell just entering the expansion zone in the first panel and in the last panel the same cell entering the differentiation zone. For this particular cell it took approximately 6 h 45 min to travel through the expansion zone. Six clones located in six roots were followed through the expansion zone, and it took 6–8 h for these clones to travel through the expansion zone. **h**, Snapshots from a video recording the growth of wild-type root in the presence of 30 nM IAA. The cells entering the expansion zone (arrows in left panel) were traced in the video to record the time it takes to enter the differentiation zone (arrows in the right panel). For the marked cell it took approximately 7 h 10 min to travel through the expansion zone. Tracking of cells through the expansion zone was carried out for nine cells located in three different roots, and it took 6–8 h for these cells to travel through the expansion zone.

32. Lee, J. Y. *et al.* Transcriptional and posttranscriptional regulation of transcription factor expression in *Arabidopsis* roots. *Proc. Natl Acad. Sci. USA* **103**, 6055–6060 (2006).



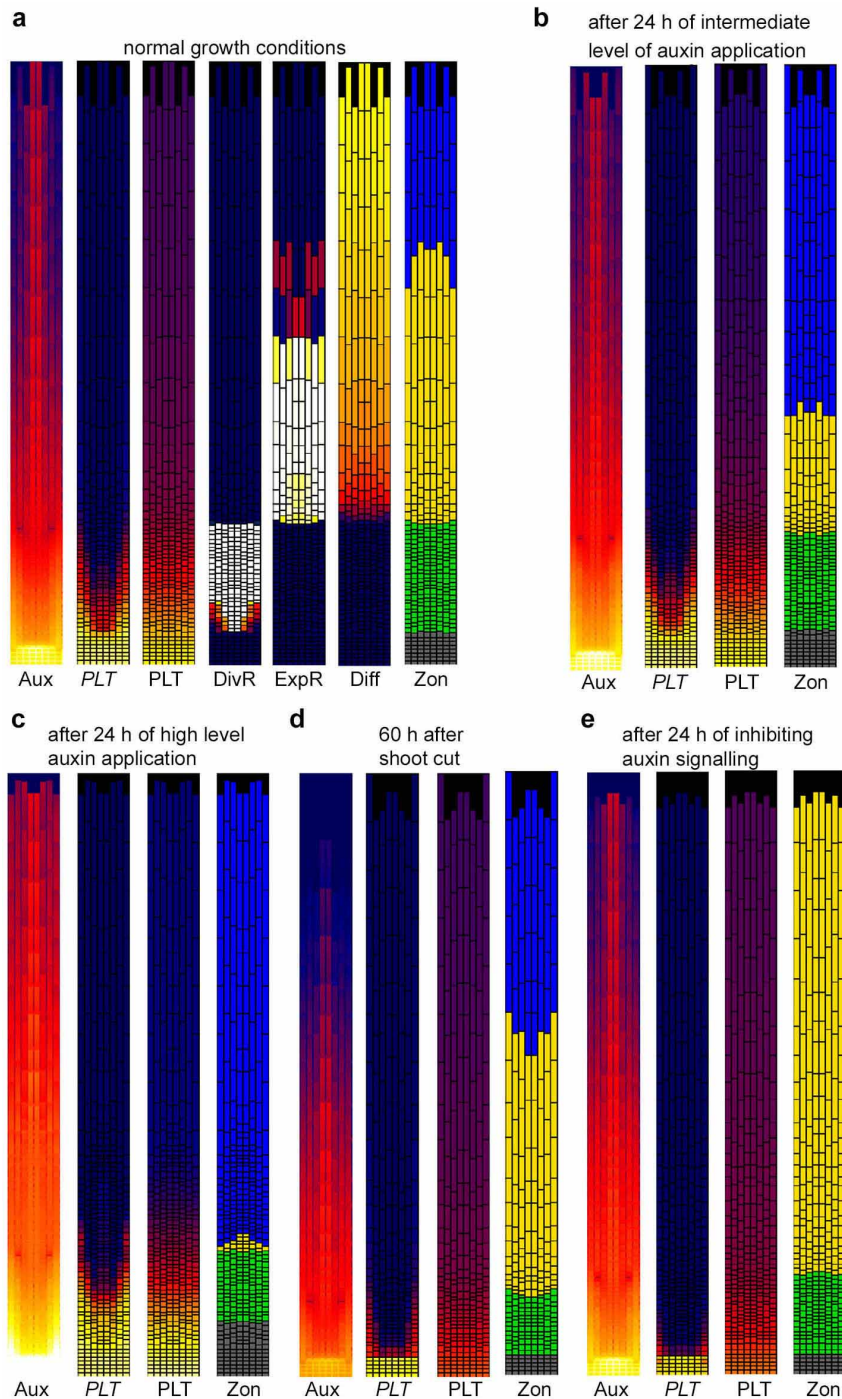


**Extended Data Figure 7 | Auxin is required for cell division, expansion and differentiation.**

**a**, Protoxylem differentiation is inhibited in the *axr3-1* mutant. First protoxylem is highlighted with an arrow. Higher up in the root protoxylem differentiation is often sporadic (inset, arrows indicate a stretch of a single protoxylem element). **b**, Twenty-four hour induction of *pG1090:XVE*  $\gg$  *axr3-1-RFP* with 17 $\beta$ -oestradiol (est) inhibits xylem differentiation and root hair outgrowth. Left inset, a single protoxylem vessel highlighted with two arrows. Right inset, the arrow highlights the beginning of a more continuous protoxylem strand, which was probably already present before the induction of *axr3-1-RFP*. Yellow bars in **b**, **c** show the areas in the root in which visible protoxylem are present. Yellow arrow in **b**, **c** marks the first root hair. Est, 5  $\mu$ M 17 $\beta$ -oestradiol. **c**, Twenty-hour treatment of 4-day old wild-type root (Col) with 25  $\mu$ M auxinole (auxin antagonist) inhibits xylem differentiation, cell expansion and root hair outgrowth. Xylem is typically differentiated as short, sporadic stretches. Left inset, arrow shows the end of a stretch of a xylem strand comprising approximately six protoxylem elements; right inset, higher up in the root two continuous protoxylem strands appear (arrows). These strands were probably already present before auxinole application. The middle inset shows the presence of short cells (marked with arrowheads) high up in the root, indicating that cell expansion was defective. The cell length typically varied along the root. **d**, Bar plot shows that the final length of cortex cells is shorter when auxin signalling is inhibited by auxinole ( $n = 35$  cells in 6 roots) or inducible *pG1090:XVE*  $\gg$  *axr3-1-RFP* ( $n = 47$  cells in

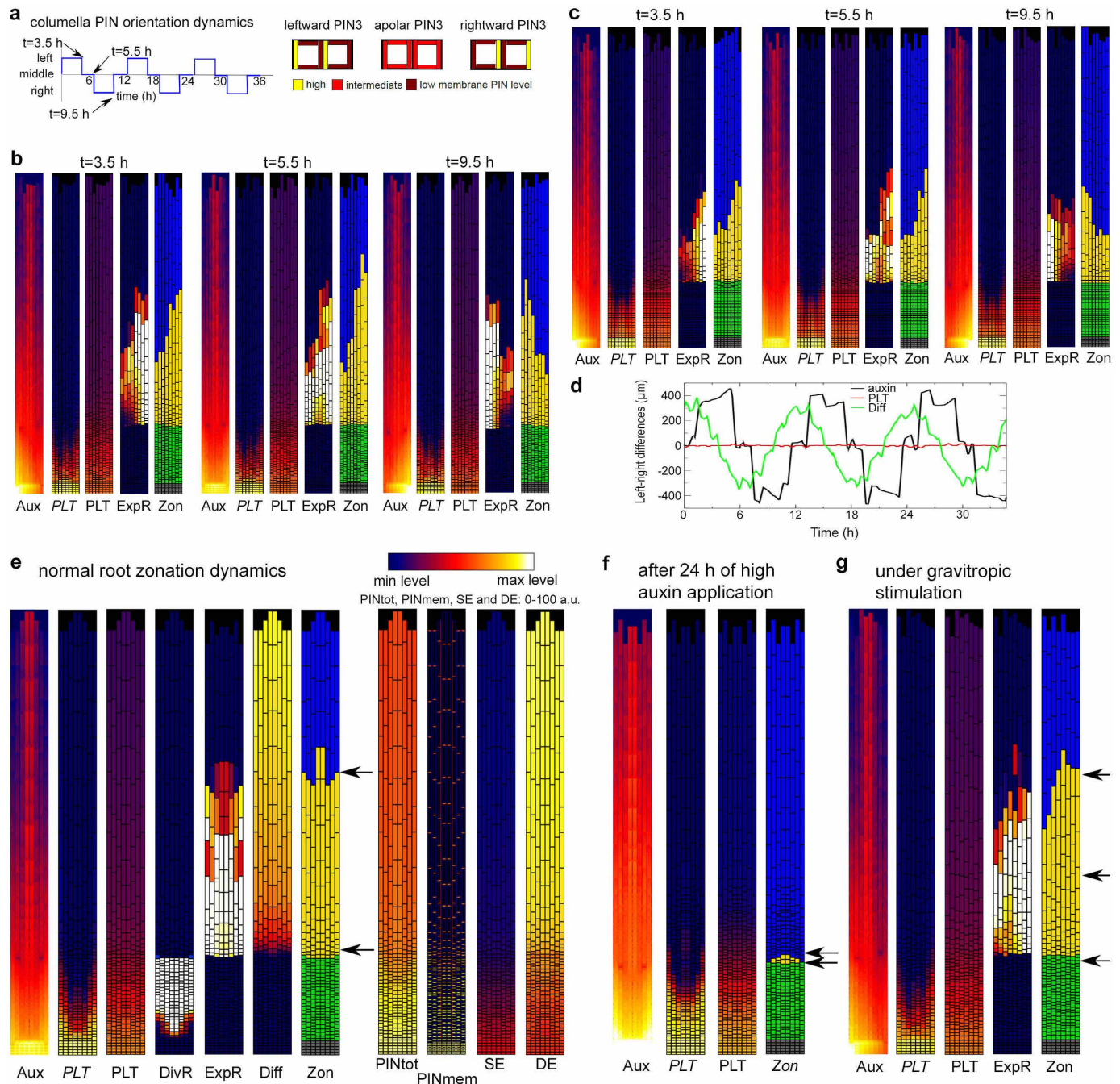
6 roots) when compared with differentiation zone cells in the control roots ( $n = 51$  cells in 7 roots and  $n = 38$  cells in 7 roots, respectively) located at a similar distance from the root tip. Asterisks, Mann–Whitney U test ( $P < 0.001$ ). Error bars show s.d. **e–j**, The consequence of auxin application or inhibition of auxin signalling on marker gene expression. **e**, **f**, Auxin application (5  $\mu$ M IAA) rapidly leads to expression of the root hair differentiation marker *pEXP7:GUS* (ref. 33) (**e**) and the xylem differentiation marker S18 (ref. 32) (**f**) in the elongation zone. Note: high auxin levels (such as 5  $\mu$ M IAA) are inhibitory for root hair elongation. Therefore, even though the root hair marker *pEXP7:GUS* rapidly spreads into the elongation zone, root hair elongation is less pronounced there. **e–g**, Twenty-four hour induction of *pG1090:XVE*  $\gg$  *axr3-1-RFP* (est) inhibits the expression of root hair differentiation marker *pEXP7:GUS* (**e**) and the xylem differentiation marker S18 (ref. 32) (**f**), as well as cell division marker *CYCB1;1-GUS*<sup>22</sup> (**g**). Note: both the signal intensity and the number of *CYCB1;1-GUS*-expressing cells are decreased in **g**. Twenty-four hour induction of *XVE*  $\gg$  *axr3-1* leads either to disappearance of S18 fluorescence (no expression in **f**), or S18 was present in short fragments (fragmented). **h**, **i**, Expression of *PLT2-YFP* continued to mark the shortened meristem after auxin signalling was inhibited by induction of *pG1090:XVE*  $\gg$  *axr3-1-RFP* (est) (**h**) or by treatment of the seedlings with 25  $\mu$ M auxinole (axnl) (**i**). **j**, Auxinole rapidly inhibits cell division as indicated by reduction of the number of *CYCB1;1-GFP*-expressing cells after auxinole treatment.

33. Cho, H. T. & Cosgrove, D. J. Regulation of root hair initiation and expansin gene expression in *Arabidopsis*. *Plant Cell* **14**, 3237–3253 (2002).



**Extended Data Figure 8 | Simulation of zonation dynamics in the auxin model under normal conditions and conditions of perturbed auxin (signalling).** **a**, Zonation dynamics under normal growth conditions. **b**, Zonation dynamics after 24 h of intermediate level auxin application. **c**, Zonation dynamics after 24 h of high level auxin application. **d**, Zonation

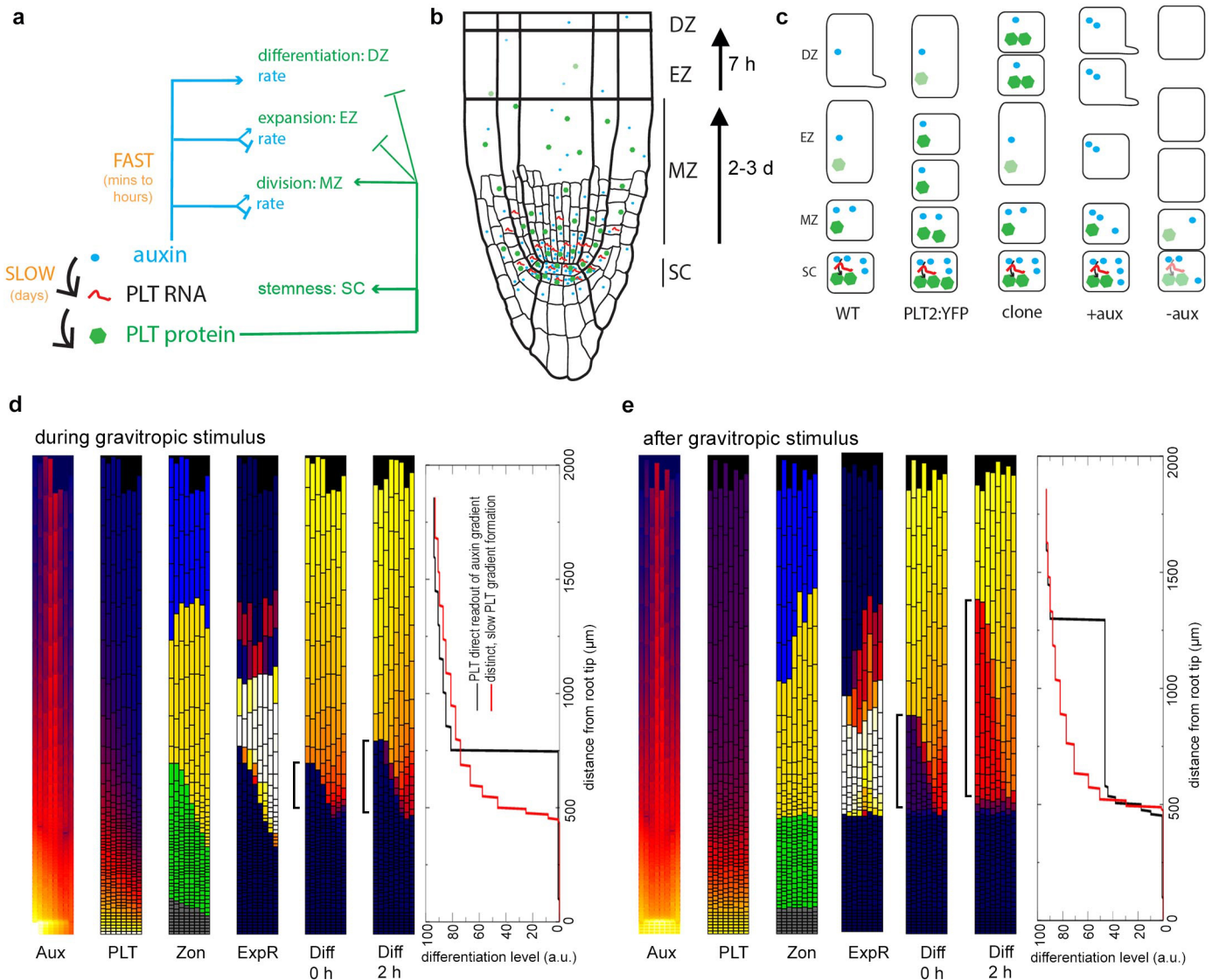
dynamics 60 h after shoot cut. **e**, Zonation dynamics after 24 h of inhibited auxin signalling. Shown are snapshots of auxin (Aux), *PLT* transcription (*PLT*), *PLT* protein (*PLT*) and zonation (*Zon*) profiles. In addition, for the normal growth conditions, snapshots of division rates (*DivR*), expansion rates (*ExpR*) and differentiation levels (*Diff*) are shown.



**Extended Data Figure 9 | Simulation of zonation under a dynamic gravistimulus protocol and in the closed feedback model.** **a**, Left, 12 h period in which leftward, apolar and rightward columella PIN orientations are interchanged to simulate dynamic gravitropism. Right, schematic depiction of the used leftward, apolar and rightward columella PIN orientations. **b**, Root zonation dynamics for the gravitropism model. Snapshots of auxin, PLT transcription, PLT protein, expansion rate and resulting zonation dynamics are shown for  $t = 3.5$  h when PIN orientation is leftward (left), at  $t = 5.5$  h when PIN orientation is apolar (middle) and at  $t = 9.5$  h when PIN orientation is rightward (right). **c**, Root zonation dynamics for the simplified gravitropism model. In the simplified gravitropism model, cellular division and differentiation rates are again constant (as in the minimal model) rather than ARF level dependent (as in the auxin and normal gravitropism models). Only expansion rates are ARF level dependent, such that they decrease from their maximum value for higher than optimal ARF levels. Similar snapshots as in **b** are shown. **d**, Dynamics of left–right differences in auxin, differentiation level and PLT protein distribution in the simplified gravitropism model.

**e–g**, Simulations with positive feedbacks from PLT back to auxin biosynthesis and transport. **e**, Zonation dynamics under standard growth conditions. In addition to the panels shown for other model versions, gene expression patterns of the genes dependent on PLT levels are shown. PINtot refers to total cellular PIN levels, PINmem to membrane PIN levels, SE to a general auxin synthesizing enzyme and DE to a general auxin degrading enzyme. Note that membrane PIN levels are a product of cellular PIN protein levels and the superimposed cell type and zone dependent membrane PIN pattern (which determines the locations and ratios of PINs deposited on the different membrane faces of the cell). **f**, Zonation dynamics after 24 h of high auxin application. **g**, Zonation dynamics under dynamic gravitropic stimulation. For comparison purposes arrows indicating the location of the transitions from MZ to EZ (bottom arrow) and from EZ to DZ (top arrow) as found in the PLT-spread model are shown. For the gravitropism simulation (**g**) the EZ to DZ transitions at both the lower (middle arrow) and upper (top arrow) side of the root are shown.





**Extended Data Figure 10 | Structure and function of the auxin-*PLETHORA* regulatory architecture.** **a**, Regulatory architecture controlling root zonation dynamics and tropisms. Slow induction of the PLTs by auxin (black arrows) defines the pathway that operates through regulating PLT levels (green arrows). Parallel to this, auxin can also control zonation rapidly without directly affecting PLT levels (blue arrows). **b**, Schematic overview of root developmental zones where local concentrations of auxin, PLT transcript and PLT proteins are represented by symbol density. **c**, Overview of the auxin, *PLT* transcript and PLT protein profiles and corresponding zonation dynamics under the following conditions: wild-type (WT), extra *PLT2* copy in wild type (*PLT2:YFP*), clonal ectopic expression of *PLT* in the expansion zone (clone), short-term auxin addition (+aux) or inhibition of auxin signalling (-aux).

Expansion is indicated by longer cell shape, differentiation by root hair bulge. **d, e**, Auxin, PLT, zonation, and expansion rate profiles during (**d**) and after (**e**) a gravitropic stimulus for a simulation in which PLT levels are a direct readout of auxin levels, and in which partly differentiated cells dedifferentiate upon re-entering the meristem. Differentiation snapshots are shown with 2 h intervals during and after the gravitropic stimulus. Brackets highlight the developmental progression of the cells that dedifferentiated under the gravitropic stimulus. Differentiation graphs show differentiation levels in the leftmost epidermal row of cells (corresponding to the side of the root towards the gravity vector) in the PLT as direct auxin readout model (black) compared to that of the model developed in this study (red), for 2 h after the (end of the) gravitropic stimulus.

# Cessation of CCL2 inhibition accelerates breast cancer metastasis by promoting angiogenesis

Laura Bonapace<sup>1,2\*</sup>, Marie-May Coissieux<sup>1\*</sup>, Jeffrey Wyckoff<sup>1†</sup>, Kirsten D. Mertz<sup>3,4</sup>, Zsuzsanna Varga<sup>3</sup>, Tobias Junt<sup>2\*</sup> & Mohamed Bentires-Alj<sup>1\*</sup>

**Secretion of C–C chemokine ligand 2 (CCL2) by mammary tumours recruits CCR2-expressing inflammatory monocytes to primary tumours and metastatic sites, and CCL2 neutralization in mice inhibits metastasis<sup>1</sup> by retaining monocytes in the bone marrow. Here we report a paradoxical effect of CCL2 in four syngeneic mouse models of metastatic breast cancer. Surprisingly, interruption of CCL2 inhibition leads to an overshoot of metastases and accelerates death. This is the result of monocyte release from the bone marrow and enhancement of cancer cell mobilization from the primary tumour, as well as blood vessel formation and increased proliferation of metastatic cells in the lungs in an interleukin (IL)-6- and vascular endothelial growth factor (VEGF)-A-dependent manner. Notably, inhibition of CCL2 and IL-6 markedly reduced metastases and increased survival of the animals. CCL2 has been implicated in various neoplasias and adopted as a therapeutic target<sup>1–3</sup>. However, our results call for caution when considering anti-CCL2 agents as monotherapy in metastatic disease and highlight the tumour microenvironment as a critical determinant of successful anti-metastatic therapy.**

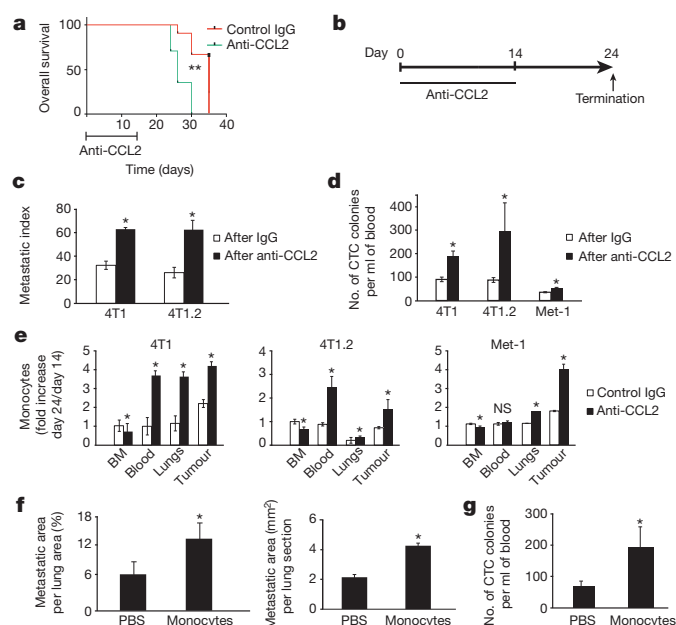
Most breast-cancer-related deaths are caused by metastases in vital organs. The tumour microenvironment is key for cancer growth, dissemination and metastasis<sup>4–6</sup>. A high number of intratumoral macrophages correlates with poor prognosis<sup>1,7–12</sup> and macrophage infiltration in breast cancer correlates with high expression of the monocyte chemoattractant CCL2. CCL2 has been proposed as a target for metastatic breast cancer because high expression of CCL2 correlates with a decrease in survival of breast cancer patients<sup>10,11,13,14</sup> (Extended Data Fig. 1a, b), and because monocytes expressing the CCL2 receptor enhance metastasis via VEGF secretion in mice<sup>1,15</sup>.

We assessed the effect of CCL2 neutralization on tumour growth and metastasis in syngeneic mouse models of CCL2-secreting metastatic breast cancer (Extended Data Fig. 1c–e). Although anti-CCL2 treatment had no effect on primary tumour growth, it reduced the number of lung metastases (Extended Data Fig. 2a–c) and circulating tumour cells (CTCs) (Extended Data Fig. 2d). Intravital imaging<sup>16</sup> showed that anti-CCL2 treatment reduced cancer cell motility and blood vessel leakiness in the tumour, and we found more pericytes around blood vessels (Extended Data Fig. 2e–g and Supplementary Videos 1, 2). Decreased blood vessel leakiness upon anti-CCL2 treatment correlated with fewer CTCs and reduced intratumoral macrophage numbers (Extended Data Fig. 3a, b). Therefore, CCL2 neutralization limits metastases not only through effects on pre-metastatic niches<sup>1</sup> but also by limiting cancer cell intravasation.

Next we examined the persistence of the anti-metastatic effect of CCL2 neutralization after treatment was discontinued. The antibody was cleared within 10 days after treatment (Extended Data Fig. 3c), leading to a rebound of CCL2 in the lungs (Extended Data Fig. 1e). Surprisingly, we found that cessation of anti-CCL2 treatment accelerated death (Fig. 1a). Ten days after interruption of anti-CCL2 treatment, we found a dramatic increase in lung and liver metastases and an increase in CTC

numbers (Fig. 1b–d and Extended Data Fig. 4a, b). Thus, although anti-CCL2 treatment reduced metastases, interruption of the treatment aggravated metastasis when compared to controls. However, the anti-metastatic effect persisted when animals were treated continuously with anti-CCL2 antibody (Extended Data Fig. 4c, d).

Since Ly6C-positive monocytes respond to CCL2, we assessed this paradoxical effect of CCL2 by focusing on the chemoattraction between breast cancer cells and SSC<sup>low</sup>CD11b<sup>+</sup>Ly6C<sup>+</sup> monocytes. We found that CCL2 drives mutual attraction of monocytes and tumour cells *in vitro* (Extended Data Fig. 5a–c). Monocytes were sequestered in the bone marrow during anti-CCL2 treatment (Extended Data Fig. 6a, b, left).



**Figure 1 | Discontinuation of anti-CCL2 treatment increases lung metastases and accelerates death of mice.** **a**, Kaplan–Meier survival curves of mice treated with anti-CCL2 antibody or control IgG for 14 days. The primary tumour was removed on the last day of treatment and the animals were left untreated until first signs of distress.  $n = 12$  mice per group (pooled data of two experiments); black box, one animal was censored.  $^{**}P = 0.0001$ , log-rank test. **b**, Timeline of the *in vivo* experiments. **c**, Metastatic index in tumour-bearing mice on day 24. **d**, Number of colonies formed by CTCs per ml blood on day 24. **e**, Ratio of SSC<sup>low</sup>CD11b<sup>+</sup>Ly6C<sup>+</sup> monocytes in organs on day 24 and on the last day of treatment (day 14). BM, bone marrow. **f**, Lung metastases were quantified as percentage of metastatic area per lung area (left) or as area per lung section (right) on five random sections per animal. **g**, Number of colonies formed by CTCs per ml blood. **c–g**, Data are shown as means  $\pm$  standard error of the mean (s.e.m.) of  $n = 8$  mice per group, pooled data from two experiments.  $^{*}P < 0.05$ , unpaired *t*-test. NS, not significant.

<sup>1</sup>Friedrich Miescher Institute for Biomedical Research (FMI), Basel 4058, Switzerland. <sup>2</sup>Novartis Institutes for Biomedical Research, 4002 Basel, Switzerland. <sup>3</sup>Department of Pathology, University Hospital Zurich, 8006 Zurich, Switzerland. <sup>4</sup>Institute of Pathology Liestal, Cantonal Hospital Baselland, 4410 Liestal, Switzerland. <sup>†</sup>Present address: Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA.

\*These authors contributed equally to this work.

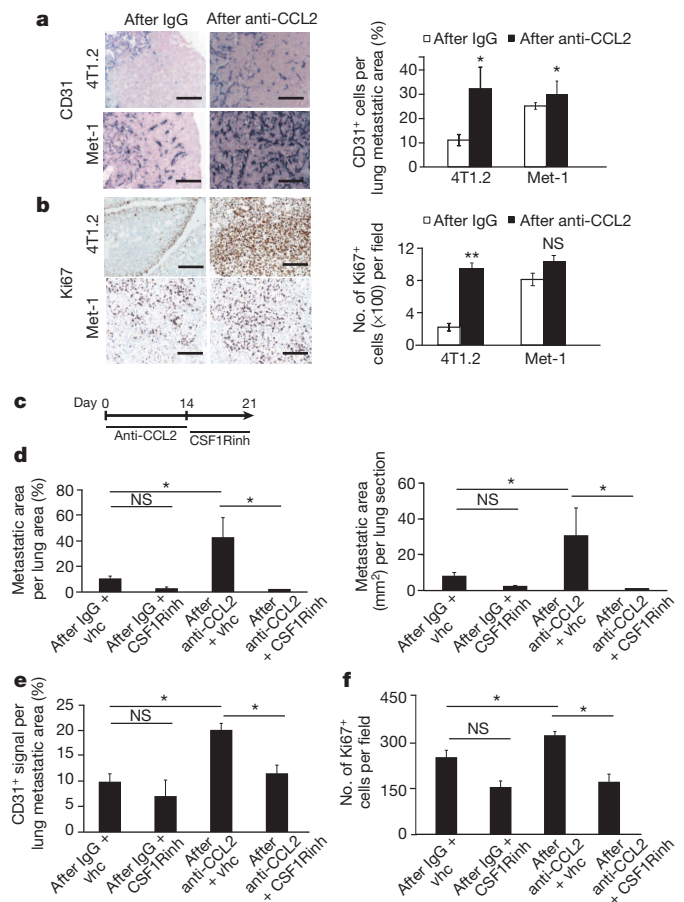
Ten days after treatment, monocyte numbers decreased in the bone marrow of animals treated with anti-CCL2 but increased in the primary tumour and lungs (Fig. 1e and Extended Data Fig. 6a–c, right). Therefore, monocytes were retained in the bone marrow during anti-CCL2 treatment and were released after treatment.

To assess monocyte distribution in tumours and metastases during and after CCL2 neutralization, we transferred sorted monocytes from bone marrow of 4T1.2-tumour-bearing mice into syngeneic tumour-bearing mice that had been treated for 7 days with anti-CCL2 or control immunoglobulin (Ig)G. During anti-CCL2 treatment, the transferred monocytes were retained in the bloodstream and their homing to the primary tumour and to the metastatic site was inhibited (Extended Data Fig. 7a). By contrast, 7 days after anti-CCL2 treatment cessation, transferred monocytes migrated to the lungs (Extended Data Fig. 7b).

To assess whether migration of monocytes into the primary tumour and lungs drives metastasis, tumour-bearing animals were transfused with monocytes from syngeneic tumour-bearing donors for 4 consecutive days. This enhanced lung metastases and numbers of CTCs (Fig. 1f, g) compared with controls. The data show that monocytes homed from the blood to the primary tumour or to the lungs in a CCL2-dependent manner and thus increased CTCs and lung metastases.

We then asked whether monocytes mediate the metastatic overshoot after cessation of anti-CCL2 treatment via enhanced mobilization of cancer cells from the primary tumour or through direct promotion of metastatic growth in the lungs. We removed the primary tumour on the last day of anti-CCL2 treatment and assessed lung metastases 10 days later. Lung metastases were still increased even though CTCs were reduced (Extended Data Fig. 8a, b). Furthermore, the number of monocytes/macrophages within lung metastases roughly doubled but increased to a lesser extent in the metastases of controls (Extended Data Fig. 8c). This suggested that the growth-enhancing effect of monocytes in the lungs crucially contributed to metastatic overshoot. Indeed, monocytes augmented the proliferation of breast cancer cells *in vitro* (Extended Data Fig. 8d). Since this direct effect was only moderate, we surmised that monocytes may also affect metastatic growth through effects on the metastatic microenvironment. We detected increased intrametastatic vasculature and Ki67<sup>+</sup> proliferating cells in metastases after anti-CCL2 discontinuation (Fig. 2a, b). Therefore, the enhancement of metastasis by monocytes could be due to a growth-enhancing effect of monocytes on tumour cells and to increased vascularization of metastases. To further test this hypothesis, we treated mice with anti-CCL2 or IgG for 14 days, followed by colony stimulating factor 1 receptor (CSF1R) inhibitor or vehicle (Fig. 2c). Since treatment with CSF1R inhibitor depletes CSF1R-positive cells<sup>17</sup> (Extended Data Fig. 8e), we concluded that monocyte depletion after anti-CCL2 treatment prevented metastatic overshoot (Fig. 2d). Vascular density and proliferation of metastases were also reduced (Fig. 2e, f and Extended Data Fig. 8e). High CCL2 in human breast cancer biopsies also correlated with high vascularization (Extended Data Fig. 9a) and with a high frequency of CD68<sup>+</sup> macrophages (Extended Data Fig. 1a). We conclude that an enhanced number of monocytes/macrophages in the lung after anti-CCL2 treatment supports local vascular growth and proliferation of metastases.

To delineate the molecular mechanism by which monocytes increase vascular density within lung metastases after anti-CCL2 treatment, we analysed cytokines in supernatants of mono- and co-cultures of cancer cells and bone marrow monocytes. Among the cytokines that were increased in tumour-monocyte co-cultures compared with monocultures (Extended Data Fig. 9b, c), IL-6 was the only one known to stimulate angiogenesis<sup>18,19</sup>. Although IL-6 was not elevated in the primary tumour after anti-CCL2, its expression in serum and metastatic lungs was higher than in controls (Extended Data Fig. 9d and Fig. 3a). In addition, IL-6 induced expression of pro-angiogenic VEGF-A in monocytes *ex vivo*. This effect appeared to be CCL2 dependent (Fig. 3b). Consistently, analysis of gene expression data sets for metastatic breast cancer revealed that concomitant high expression of IL-6 in tumours with high CCL2 expression was associated with diminished patient survival, compared

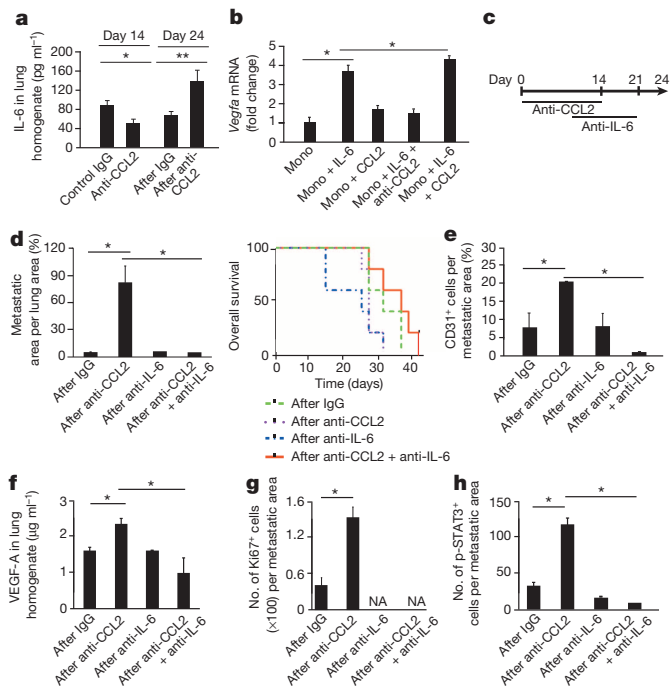


**Figure 2 | Monocytes induce blood vessel formation and proliferation of lung metastases after discontinuation of anti-CCL2 treatment.** **a**, Left, representative images of CD31-stained lung metastases on day 24. Scale bar, 50  $\mu$ m. Right, quantification of CD31 staining. **b**, Left, representative images of Ki67-stained lung metastases on day 24. Scale bar, 100  $\mu$ m. Right, quantification of Ki67 staining. **a**, **b**, Graphs show means  $\pm$  s.e.m. of 20 fields of view on 5 lung sections per animal,  $n = 4$  mice per group, pooled data from two experiments.  $^{**}P < 0.001$ ,  $^{*}P < 0.05$ , unpaired *t*-test. One representative image out of 20 is shown per group. **c**, Timeline of *in vivo* experiments. CSF1Rinh, CSF1R inhibitor. **d**, Lung metastases were quantified as percentage of metastatic area per lung area (left) or as area per lung section (right). vhc, vehicle. **e**, Quantification of CD31 staining. **f**, Quantification of Ki67 staining. **d–f**, Graphs show means  $\pm$  s.e.m. of 10 fields of view on 5 lung sections per animal,  $n = 4$  mice per group, pooled data from two experiments.  $^{*}P < 0.05$ , analysis of variance (ANOVA) with post-hoc Bonferroni correction. NS, not significant.

with concomitant low expression of IL-6 (Extended Data Fig. 9e). We conclude that high expression of IL-6 in the lungs subsequent to the cessation of anti-CCL2 treatment induces VEGF-A in monocytes and, at the same time, increases local vascular density.

Since IL-6 appeared to drive blood vessel density in metastatic lungs, we treated mice with anti-CCL2 followed by anti-IL-6. We observed reduced lung metastases and CTCs and increased survival of the mice (Fig. 3c, d and Extended Data Fig. 9f). We then asked whether anti-IL-6 decreased angiogenesis and tumour cell proliferation within metastases. Treatment of tumour-bearing animals with anti-IL-6 after anti-CCL2 reduced the overshooting number of CD31<sup>+</sup> cells in lungs and VEGF-A levels in lungs and serum (Fig. 3e, f and Extended Data Fig. 9g). Consistently, the proliferation of metastatic tumour cells declined to undetectable levels (Fig. 3g). Next we sought to understand how IL-6 induces VEGF-A expression *in vivo*. STAT3 phosphorylation (p-STAT3) results in VEGF-A secretion downstream of IL-6 signalling<sup>20,21</sup>. Therefore we assessed the level of p-STAT3 in lung metastatic tissue. Consistent with



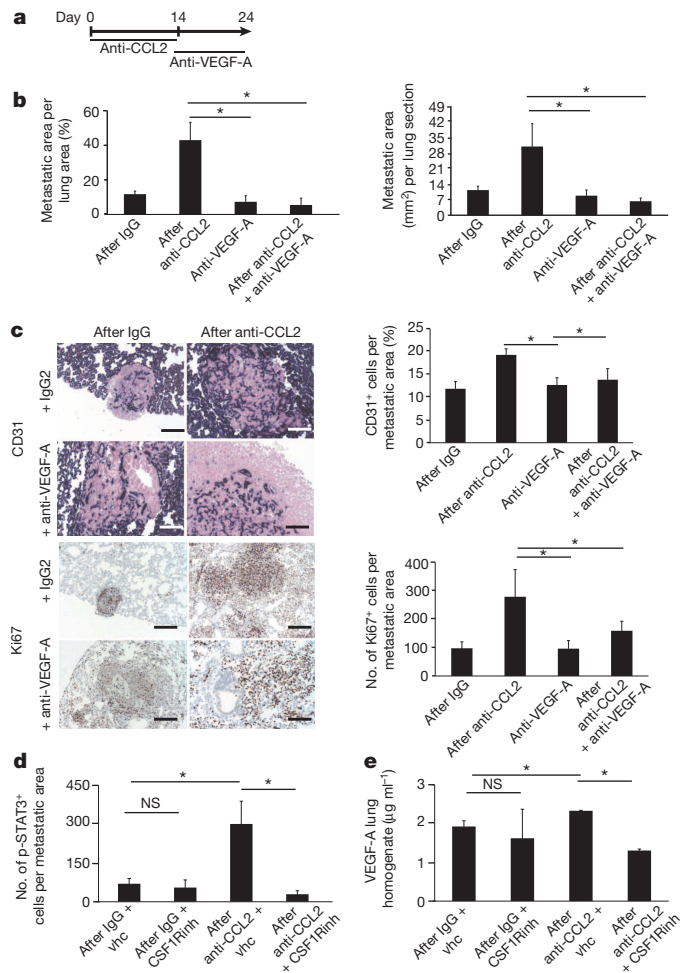


**Figure 3 | Anti-IL-6 treatment after cessation of anti-CCL2 therapy prevents the overshoot of lung metastases.** **a**, IL-6 expression in lung homogenates on day 14 and 10 days after anti-CCL2 treatment. Data are shown as means  $\pm$  s.e.m.,  $n = 4$  mice per group, pooled data from two experiments.  $*P < 0.05$ ,  $**P < 0.001$ , unpaired *t*-test. **b**, Mean *Vegfa* messenger RNA levels in monocytes (Mono) from tumour-bearing mice treated *ex vivo* with IL-6 and/or anti-CCL2. Data show means  $\pm$  s.e.m.,  $n = 3$  mice per group, one representative of three independent experiments.  $*P < 0.05$ , ANOVA with post-hoc Bonferroni correction. **c**, Timeline of *in vivo* experiments. **d**, Left, quantification of lung metastases. Right, Kaplan-Meier survival curves of animals treated with control IgG, anti-CCL2, anti-IL-6 or anti-CCL2 plus anti-IL-6 as in Fig. 3c. On day 24, the animals were left without treatment until the appearance of signs of distress.  $n = 10$  mice per group.  $P = 0.0017$  by the log-rank test comparing anti-CCL2 with anti-CCL2 plus anti-IL-6 treatment. **e**, Quantification of CD31 staining. **f**, VEGF-A expression in lung homogenates from mice treated as in Fig. 3c measured by enzyme-linked immunosorbent assay (ELISA). Data are shown as means  $\pm$  s.e.m.,  $n = 8$  mice per group, pooled data from two experiments.  $*P < 0.05$ , ANOVA with post-hoc Bonferroni correction. **g**, Quantification of Ki67 staining. **d** (left), **e**, **g**, Data show means  $\pm$  s.e.m. of 20 fields of view on 5 sections per animal,  $n = 4$  animals per group, pooled data from two experiments.  $*P < 0.05$ , ANOVA with post-hoc Bonferroni correction. NA, not applicable. **h**, Quantification of p-STAT3 staining. Data are shown as means  $\pm$  s.e.m. of 10 fields of view on 5 sections per animal,  $n = 4$  mice per group, pooled data from two experiments.  $*P < 0.05$ , ANOVA with post-hoc Bonferroni correction.

the increases in IL-6 and VEGF-A in the lungs, p-STAT3 also increased upon discontinuation of anti-CCL2 (Fig. 3h). These results suggest that IL-6/p-STAT3 pro-angiogenic effects contributed to the metastatic overshoot after cessation of anti-CCL2 treatment.

We neutralized VEGF-A *in vivo* to assess whether it supported tumour cell proliferation after anti-CCL2 or IgG treatment (Fig. 4a). This prevented metastatic overshoot, reduced blood-vessel density and reduced proliferation in lung metastases (Fig. 4b, c). p-STAT3 levels in the lungs did not change with this treatment (Extended Data Fig. 10a), suggesting that VEGF-A acts downstream of p-STAT3.

To confirm that monocytes/macrophages contributed to the expression of VEGF-A after cessation of anti-CCL2, we assessed p-STAT3 and VEGF-A levels in lungs from animals treated with anti-CCL2 or IgG control, followed by macrophage depletion with CSF1R inhibitor. p-STAT3 and VEGF-A levels increased in lungs of animals after cessation of anti-CCL2 treatment, but not if anti-CCL2 treatment was followed by monocyte/macrophage depletion (Fig. 4d, e and Extended Data Fig. 10b).



**Figure 4 | Combined anti-CCL2 and anti-VEGF treatment reduces angiogenesis and tumour cell proliferation in lung metastases.** **a**, Timeline of the *in vivo* experiments. **b**, Quantification of lung metastases as percentage of metastatic area per lung area (left) or as area per lung section (right). Data are shown as means  $\pm$  s.e.m. of 20 fields of view on 5 lung sections per animal,  $n = 8$  mice per group, pooled data from two experiments.  $*P < 0.05$ , ANOVA with post-hoc Bonferroni correction. **c**, Top left, representative images of CD31-stained lungs on day 24. Top right, quantification of CD31 staining. Data are shown as means  $\pm$  s.e.m. of 20 fields of view on 5 lung sections per animal,  $n = 4$  mice per group, pooled data from two experiments.  $*P < 0.05$ , ANOVA with post-hoc Bonferroni correction. Scale bar, 100  $\mu$ m. Bottom left, representative images of Ki67-stained lung on day 24. Bottom right, quantification of Ki67 staining. Data are shown as means  $\pm$  s.e.m. of 20 fields of view on 5 lung sections per animal,  $n = 3$  mice per group, pooled data from two experiments.  $*P < 0.05$ , ANOVA with post-hoc Bonferroni correction. Scale bar, 100  $\mu$ m. One representative image out of 20 is shown per group. **d**, Quantification of p-STAT3 staining. Data are shown as means  $\pm$  s.e.m. of 10 fields of view on 5 sections per animal,  $n = 3$  mice per group, pooled data from two experiments.  $*P < 0.05$ , ANOVA with post-hoc Bonferroni correction. CSF1Rinh, CSF1R inhibitor; vhc, vehicle. **e**, VEGF-A expression in lung homogenates from mice treated as in Fig. 4a measured by ELISA. Data are shown as means  $\pm$  s.e.m.,  $n = 3$  mice per group, pooled data from two experiments.  $*P < 0.05$ , ANOVA with post-hoc Bonferroni correction. NS, not significant.

This suggested that IL-6/p-STAT3-dependent VEGF-A secretion in the lung after cessation of CCL2 neutralization is indeed mediated by lung macrophages.

Together, these data support the idea that the complex tumour micro-environment is critical for effective anti-tumour strategies. Anti-CCL2 treatment decreased breast cancer metastases in mice, but interruption of anti-CCL2 treatment precipitated an unexpected influx of monocytes into the metastatic site and overshooting IL-6 levels within the

metastatic microenvironment. This led to local enhancement of angiogenesis, metastatic disease and a fatal outcome. Therefore our results prompt extreme caution when considering anti-CCL2 treatment of metastatic breast cancer and other neoplasias and suggest that therapeutic interference with the tumour microenvironment might lead to tissue remodelling not only locally but also at remote sites such as the bone marrow, with unexpected far-reaching consequences including a worsened prognosis for cancer patients. Any tumour immunotherapy that only sequesters immune cells away from the tumour and that does not permanently reprogram the tissue microenvironment or directly kill tumour cells may bear a similar risk of lethal rebound.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 1 December 2013; accepted 15 September 2014.**

**Published online 22 October 2014.**

- Qian, B. Z. *et al.* CCL2 recruits inflammatory monocytes to facilitate breast-tumour metastasis. *Nature* **475**, 222–225 (2011).
- Lu, X. & Kang, Y. Chemokine (C-C motif) ligand 2 engages CCR2<sup>+</sup> stromal cells of monocytic origin to promote breast cancer metastasis to lung and bone. *J. Biol. Chem.* **284**, 29087–29096 (2009).
- Wolf, M. J. *et al.* Endothelial CCR2 signaling induced by colon carcinoma cells enables extravasation via the JAK2-Stat5 and p38MAPK pathway. *Cancer Cell* **22**, 91–105 (2012).
- Pietras, K. & Ostman, A. Hallmarks of cancer: interactions with the tumor stroma. *Exp. Cell Res.* **316**, 1324–1331 (2010).
- McAllister, S. S. & Weinberg, R. A. Tumor-host interactions: a far-reaching relationship. *J. Clin. Oncol.* **28**, 4022–4028 (2010).
- Hanahan, D. & Coussens, L. M. Accessories to the crime: functions of cells recruited to the tumor microenvironment. *Cancer Cell* **21**, 309–322 (2012).
- Kamoub, A. E. & Weinberg, R. A. Chemokine networks and breast cancer metastasis. *Breast Dis.* **26**, 75–85 (2007).
- Kleer, C. G., van Golen, K. L. & Merajver, S. D. Molecular biology of breast cancer metastasis. Inflammatory breast cancer: clinical syndrome and molecular determinants. *Breast Cancer Res.* **2**, 423–429 (2000).
- Palangie, T. *et al.* Prognostic factors in inflammatory breast cancer and therapeutic implications. *Eur. J. Cancer* **30**, 921–927 (1994).
- Ueno, T. *et al.* Significance of macrophage chemoattractant protein-1 in macrophage recruitment, angiogenesis, and survival in human breast cancer. *Clin. Cancer Res.* **6**, 3282–3289 (2000).
- Valković, T., Lucin, K., Krstulja, M., Dobi-Babic, R. & Jonjic, N. Expression of monocyte chemoattractant protein-1 in human invasive ductal breast cancer. *Pathol. Res. Pract.* **194**, 335–340 (1998).
- Saji, H. *et al.* Significant correlation of monocyte chemoattractant protein-1 expression with neovascularization and progression of breast carcinoma. *Cancer* **92**, 1085–1091 (2001).
- Chavey, C. *et al.* Oestrogen receptor negative breast cancers exhibit high cytokine content. *Breast Cancer Res.* **9**, R15 (2007).
- Goede, V., Brogelli, L., Ziche, M. & Augustin, H. G. Induction of inflammatory angiogenesis by monocyte chemoattractant protein-1. *Int. J. Cancer* **82**, 765–770 (1999).
- Li, X. *et al.* A destructive cascade mediated by CCL2 facilitates prostate cancer growth in bone. *Cancer Res.* **69**, 1685–1692 (2009).
- Bonapace, L. *et al.* If you don't look, you won't see: intravital multiphoton imaging of primary and metastatic breast cancer. *J. Mammary Gland Biol. Neoplasia* **17**, 125–129 (2012).
- Strachan, D. C. *et al.* CSF1R inhibition delays cervical and mammary tumor growth in murine models by attenuating the turnover of tumor-associated macrophages and enhancing infiltration by CD8 T cells. *Oncol. Immunology* **2**, e26968 (2013).
- Nilsson, M. B., Langley, R. R. & Fidler, I. J. Interleukin-6, secreted by human ovarian carcinoma cells, is a potent proangiogenic cytokine. *Cancer Res.* **65**, 10794–10800 (2005).
- Wani, A. A., Jafarnejad, S. M., Zhou, J. & Li, G. Integrin-linked kinase regulates melanoma angiogenesis by activating NF- $\kappa$ B/interleukin-6 signaling pathway. *Oncogene* **30**, 2778–2788 (2011).
- Wei, D. *et al.* Stat3 activation regulates the expression of vascular endothelial growth factor and human pancreatic cancer angiogenesis and metastasis. *Oncogene* **22**, 319–329 (2003).
- Wei, L. H. *et al.* Interleukin-6 promotes cervical tumor growth by VEGF-dependent angiogenesis via a STAT3 pathway. *Oncogene* **22**, 1517–1527 (2003).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** We thank R. L. Anderson, N. Hynes, and R. Cardiff for cell lines, R. Thierry and M. Kirschmann for scripts for immunohistochemistry and two-photon imaging analysis, M. Stadler and H.-R. Hotz for bioinformatics, H. Kohler for FACS and S. Bichet for technical support, H. Brinkhaus and P. Loetscher for helpful discussions, D. Dylan for the CSFR1 inhibitor, J. van Rheenen, J. Stein, J. Rietdorf, T. Oertner and S. Bundschuh for helping us set up the multiphoton intravital microscope, and the Bentires-Alj group for feedback. Research in the laboratory of M.B.-A. is supported by the Novartis Research Foundation, the European Research Council (243211-PTPBD), the Swiss Cancer League, the Swiss National Foundation, and the Krebsliga Beider Basel. M.-M.C. is supported by the FP7 Marie Curie Fellowship.

**Author Contributions** L.B., M.-M.C., T.J. and M.B.-A. designed and performed most of the experiments and wrote the manuscript. K.D.M. and Z.V. performed human biopsy experiments. J.W., M.-M.C. and L.B. performed intravital imaging. All authors prepared the manuscript.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.B.-A. (bentires@fmi.ch).

## METHODS

**Patient cohorts.** Samples from paraffin-embedded invasive breast cancer tissues ( $n = 30$ ) were retrieved from the archives of the Institute of Surgical Pathology, University Hospital Zurich and the Institute of Pathology, Cantonal Hospital Baselland, Liestal, Switzerland. The minimum follow-up time was 2 years. Clinicopathological data for all patients were available. All biopsies were collected for previous diagnostic purposes and processed according to standard procedures. Two board-certified pathologists (K.D.M., Z.V.) verified the diagnoses of invasive breast carcinoma.

The retrospective study on formalin-fixed, paraffin-embedded human breast cancer tissue was approved by the Cantonal Ethical Committee of Zurich (KEK-2012-553). Informed consent was not necessary, as the ethical approval covered the ethical issues of the retrospective study and the samples were completely anonymized and de-identified before the study. For immunohistochemical studies, two groups of breast cancer patients were examined: Group 1 ( $n = 14$ ; 12 hormone-receptor positive, 2 triple negative) were primary breast cancer patients without lymph node or distal metastases; Group 2 patients ( $n = 16$ ; 14 hormone-receptor positive, 2 triple negative) had both lymph node involvement and distant metastases, except for three patients with lymph node metastases whose distant metastasis status was unknown at the time of this study.

**Immunohistochemistry.** For all cases, three sequential whole-tissue sections were stained with the antibodies anti-CD68 (DAKO-Cytomation, clone PG-M1, 1:100 dilution), anti-CD31 (Novocastra, clone 1A10, 1:100 dilution, Leica Biosystems) and anti-CCL2 (R&D Systems, clone 23002, 1:100 dilution). Immunohistochemistry on 4- $\mu$ m sections was performed using an automated immunohistochemistry Bond platform (Leica Biosystems). Detection was carried out with Refine-Red or Refine-DAB kits (Bond), including the respective secondary antibodies. The sections were visualized with an Olympus BX41 microscope and attached Olympus UC30 camera using the cell<sup>A</sup> digital imaging acquisition software (Olympus).

The number of CD68-positive macrophages per high-power field (HPF) was counted automatically. CD68-positive elements were only counted as macrophages if a nucleus could be identified in the section and only intratumoral macrophages were taken into account. The mean of 20 HPFs was calculated for each case. Similarly, the number of CD31-positive vessels per HPF was counted automatically (20 HPFs per case). Only areas of positive staining with a minimal diameter of 20  $\mu$ m were taken into account to exclude isolated CD31<sup>+</sup> cells. Cytoplasmic CCL2 staining within tumour cells was scored manually by a board-certified pathologist (K.D.M.) according to a three-tiered system: 0 (negative staining, no detectable staining), 1+ (staining in <10% of tumour cells), 2+ (staining in >10% of tumour cells).

Tumours grown in mice were fixed in 10% neutral buffered formalin (NBF) for 24 h at 4 °C, washed with 70% ethanol, embedded in paraffin, sectioned at 3  $\mu$ m, and stained with haematoxylin and eosin (H&E), anti-Ki67 (clone PA5-19462, 1:400 dilution, Thermo Scientific) or anti-CD31 (clone 390, 1:500 dilution, Invitrogen) antibodies. The investigator was blinded in all the immunohistochemistry quantifications.

**Analysis of public microarray data.** The publicly available processed data sets from The Cancer Genome Atlas (TCGA) for breast invasive carcinoma (AgilentG4502A, version 2013-07-12) were downloaded from <https://genome-cancer.ucsc.edu/> and analysed with R/bioconductor<sup>22–25</sup>. Patients (all clinical subtypes) were split according to their CCL2 and IL-6 expression and the highest tertile were compared to the lowest tertile for each gene. The resulting groups were tested for difference in overall survival using the survival package in R by fitting a Cox proportional hazard regression model.

**Reagents, recombinant proteins and antibodies.** Recombinant mouse JE/MCP-1 (CCL2) and mouse IL-6 were purchased from Peprotech (Switzerland). Rat anti-mouse-IL-6 (clone MAB-406) and rat IgG1 control were from R&D Systems (UK). Purified NA/LE hamster anti-mouse MCP-1/CCL2 (clone 2H5) and hamster IgG2a control were purchased from BD Bioscience (Switzerland). Goat anti-mouse VEGF-A (clone AF493NA) and polyclonal goat IgG (R&D systems) were used for cell culture (1  $\mu$ g ml<sup>-1</sup>) and for *in vivo* experiments. CSF1R inhibitor was provided by D. Dylan<sup>17</sup>.

The following anti-mouse antibodies were used for flow cytometry: CD11b (anti-mouse APC-CD11b, clone M1/70), Ly6C (anti-mouse FITC-Ly6C, clone HK1.4), Ly6G (anti-mouse Percp-Cy5.5-Ly6G, clone IA8), CD206 (anti-mouse FITC-CD206, clone C068C2), I-A/I-E (anti-mouse Percp-Cy5.5-I-A/I-E, clone M5/114.15.2), CD11c (anti-mouse Percp-Cy5.5-CD11c, clone N418), CD86 (anti-mouse Alexa-fluor 488 CD86, clone GL-1), all from BD Bioscience, and mCCR2 (anti-mouse FITC-CCR2), from R&D Systems.

**Cell culture.** Syngeneic mammary cancer 4T1 cells<sup>26</sup>, 4T1.2 cells (derived from Balb/c mice, from R. L. Anderson), J110 cells (derived from FVB mice, from N. Hynes) and Met-1 cells (from FVB mice, from R. Cardiff) were cultured in DMEM supplemented with 10% FCS. In co-culture experiments, 10<sup>5</sup> tumour cells were incubated with 10<sup>5</sup> sorted primary bone marrow monocytes in DMEM 0.5% FCS. Cell viability was measured using a Neubauer chamber and trypan blue.

**ELISA and cytokine arrays.** CCL2 levels were assessed using the mCCL2 ELISA Duo-Set (R&D Systems) for mouse cells and the hCCL2 ELISA Duo-Set (R&D Systems) for human cells. IL-6 was measured by mIL-6 ELISA Duo-Set (R&D Systems). VEGF-A levels were assessed using the mVEGF-A Duo-set (R&D System). Cytokine arrays from R&D Systems were used according to the manufacturer's protocol.

**Flow cytometry.** Cultured adherent cells were detached using trypsin-EDTA, suspended in growth medium and counted. Tumours, lungs and bone marrow were mechanically and enzymatically dissociated (using collagenase II, Sigma). Cells were incubated with 2.5  $\mu$ g antibodies per 10<sup>6</sup> cells or with 2.5  $\mu$ g mouse IgG isotype control antibody (R&D Systems) per 10<sup>6</sup> cells for 20 min at 4 °C in the dark before washing and analysis. At least 10<sup>4</sup> cells per sample were analysed with a FACScalibur flow cytometer (Becton Dickinson). Monocytes were identified for quantification and sorting as SSC<sup>low</sup>CD11b<sup>+</sup>Ly6C<sup>+</sup> cells. The reference population for gating was SSC<sup>low</sup>CD11b<sup>+</sup>, thereby excluding granulocytes, as shown in Extended Data Fig. 5a. Absolute monocyte numbers were quantified as previously described<sup>27</sup>. In brief, after organ harvest, single-cell suspensions were obtained from bone marrow, blood, lung, or tumour. Total viable cell numbers were determined using Trypan Blue. Absolute monocyte numbers were calculated as number of total viable cells multiplied by the percentage of SSC<sup>low</sup>CD11b<sup>+</sup>Ly6C<sup>+</sup>Gr1<sup>+</sup> monocytes of all live cells.

**RNA and quantitative real-time PCR.** Total RNA was extracted using the Trizol reagent according to the manufacturer's protocol (Ambion RNA, Life Technologies). Aliquots of 1  $\mu$ g of total RNA were transcribed using the SuperScript III First-strand synthesis for RT-PCR (Invitrogen). PCR and fluorescence detection were performed using the StepOnePlus Sequence Detection System (Applied Biosystems) according to the manufacturer's protocol in a reaction volume of 20  $\mu$ l containing 1 $\times$  TaqMan Universal PCR Master Mix (Applied Biosystems) and 30 ng cDNA. For quantification of mouse *Ccl2*, *Hprt1*, *Vegfa* and *Il6* mRNA, the 1 $\times$  TaqMan Gene Expression Assays Mm\_00441242\_m1, Mm00446968\_m1, Mm01281449\_m1 and Mm00446190\_m1 (Applied Biosystems) were used. All measurements were performed in duplicates and the arithmetic means of the cycle threshold (Ct) values were used for calculations: target gene mean Ct values were normalized to the respective housekeeping gene (*Hprt1*), mean Ct values (internal reference gene, Ct), and then to the experimental control. The values obtained were exponentiated 2 ( $-\Delta\Delta Ct$ ) to be expressed as  $n$ -fold changes in regulation compared with the experimental control (2( $-\Delta\Delta Ct$ )) by the method of relative quantification<sup>25</sup>.

**Animal experiments.** Experiments using FVB and Balb/c (Charles River Laboratory and Jackson Laboratory, respectively) mice were carried out according to Swiss national guidelines on animal welfare and the regulations of the Canton of Basel Stadt.

To induce spontaneous metastasis, 1–4  $\times$  10<sup>6</sup> tumour cells of different lines were injected into the fourth mammary fat pad of 7- to 8-week-old female FVB or Balb/c mice. Antibody treatments were performed 3–15 days after tumour cell injection (depending on the cell line used) via intraperitoneal injection 3 times per week. Anti-CCL2, anti-IL6 and anti-VEGF antibodies or the respective controls were used at 1 mg kg<sup>-1</sup>. The CFS1R inhibitor (200 mg kg<sup>-1</sup> oral dosing) and the corresponding vehicle were administered daily.

Tumour volume was calculated using the formula: length  $\times$  width<sup>2</sup>/2. To assess the number of CTCs, blood was taken from the right atrium, plated in DMEM medium supplemented with 10% FCS, and colonies counted on day 10 of culture. The number of CTCs was calculated as the total number of colonies in the dish divided by the volume of blood taken. To quantify metastases, the lungs were fixed in Bouin's solution (Sigma Aldrich) for macrometastases or in PFA 4% for micrometastases.

For adoptive transfer of monocytes, tumour-bearing animals were killed on day 14 after orthotopic injection of cancer cell lines. SSC<sup>low</sup>CD11b<sup>+</sup>Ly6C<sup>+</sup> monocytes were sorted from femoral bone marrow into PBS and 1% FCS, stained with the vital dye CMTMR (5-(and-6)-((4-chloromethyl)benzoyl)amino)tetramethylrhodamine; Invitrogen) for 10 min at 37 °C according to the manufacturer's protocol. Aliquots of 8  $\times$  10<sup>6</sup> labelled cells were injected into the tail vein of syngeneic tumour-bearing mice that were treated with anti-CCL2 or the IgG control by intraperitoneal injection of fresh antibodies diluted in PBS (10 mg kg<sup>-1</sup>). Mice were killed 1 day later and single-cell suspensions of the lungs, blood, bone marrow and primary tumours were counterstained with CD11b-APC and analysed by flow cytometry. To assess the effect of monocytes on metastases, 1  $\times$  10<sup>6</sup> sorted SSC<sup>low</sup>CD11b<sup>+</sup>Ly6C<sup>+</sup> cells from tumour-bearing animals were injected into the tail vein of recipient syngeneic tumour-bearing animals daily for 4 consecutive days. The mice were killed on day 11 after tumour cell injection and CTCs and metastases were assessed as described earlier.

**Intravital imaging.** Orthotopic mammary tumours were grown for 1–2 weeks. Mice were anaesthetized with Attane Isofluran (Provect AG) and mounted on a custom-made stage. Anaesthesia was maintained throughout the experiment with a nose cone. Tumours were exposed by skin flap surgery<sup>28</sup> on a custom-made multiphoton microscope<sup>16</sup> and imaged at 880 nm with a  $\times$ 25/1.05NA water immersion objective (Olympus) at a resolution of 1.06  $\mu$ m per pixel. Cell motility was observed by time-lapse imaging over 30 min in 2-min cycles, where a 100  $\mu$ m z-stack at



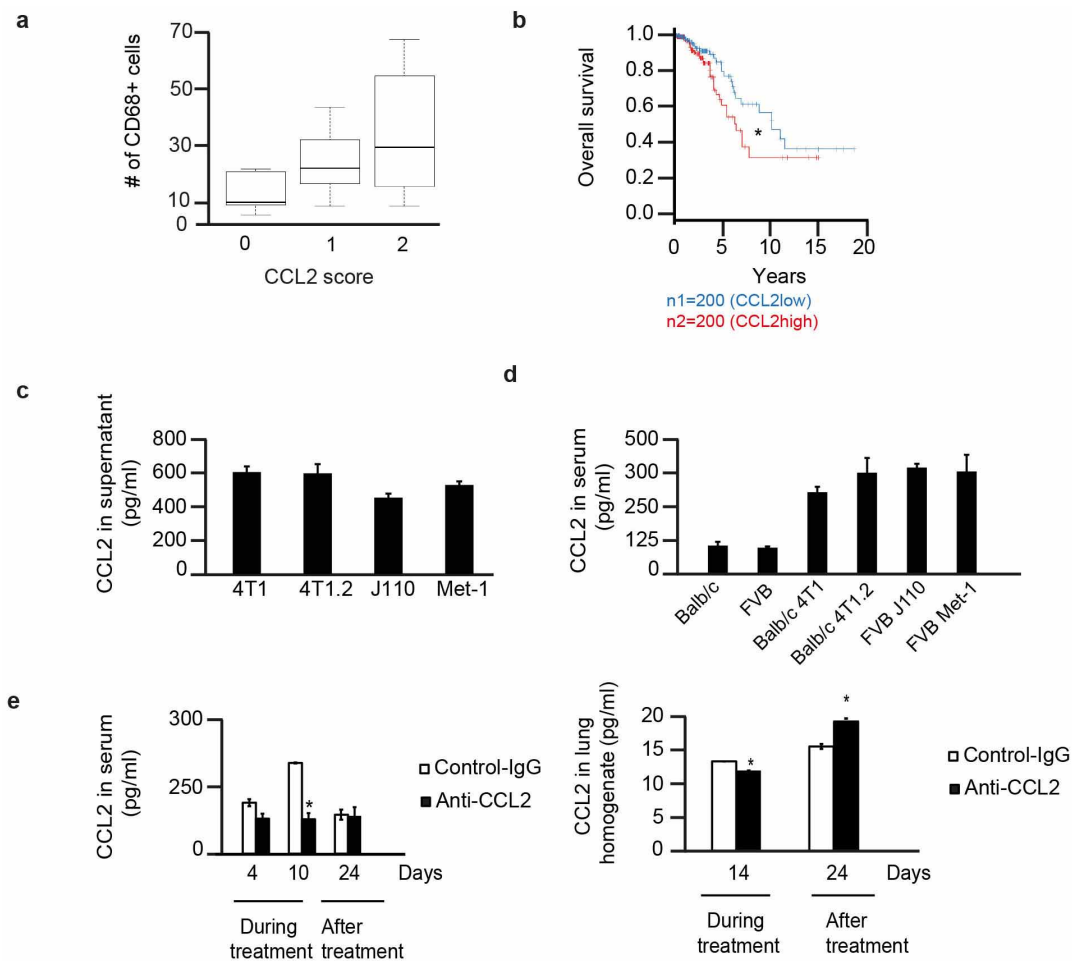
5- $\mu\text{m}$  increments was recorded for each frame starting at the tumour capsule. Three-dimensional time-lapse videos were analysed using Image J<sup>29</sup>. Tumour cell motility was quantified manually. A tumour cell motility event was defined as a protrusion of half a cell length or more over the course of a 30 min video. For visualizing vasculature, 100  $\mu\text{l}$  of 20 mg ml<sup>-1</sup> 70 kDa Texas Red-dextran (Invitrogen, Molecular Probes) was injected into the tail vein of the mice before surgery and imaging of blood vessels and macrophages. Vasculature was quantified manually as tubes of at least 20  $\mu\text{m}$  length appearing in the red channel. Permeability of the blood vessels was measured as the time taken by dextran to leak out from the blood vessels as reported previously<sup>30</sup>.

**Cell invasion.** Invasion assays were performed using transwell chambers (8  $\mu\text{m}$  pore size, BD Biocoat Growth factor-reduced Matrigel Invasion Chamber) according to the manufacturer's protocol. An aliquot of  $5 \times 10^4$  cells in 100  $\mu\text{l}$  of DMEM 0.1% BSA was plated into the top chamber and DMEM 10% FBS alone or containing monocytes derived from bone marrow of animals bearing autologous tumours was placed in the bottom chamber. After a 24 h incubation at 37 °C and 5% CO<sub>2</sub>, cells at the lower surface of the membrane were fixed with 3.7% paraformaldehyde, stained with 0.2% crystal violet, and washed with 1  $\times$  PBS. The number of cells per field was quantified microscopically and total cell number was evaluated by trypan blue vital.

**Statistical analysis.** All the *in vitro* experiments were performed in biological and technical triplicates. The number of mice was calculated by performing power analysis using data from small pilot experiments. Values represent the means  $\pm$  s.e.m. Depending on the type of experiment, data were tested using two-tailed Student's *t*-test, log-rank test, or one-way ANOVA with post-hoc Bonferroni correction. \**P* < 0.05 and \*\**P* < 0.001 were considered statistically significant.

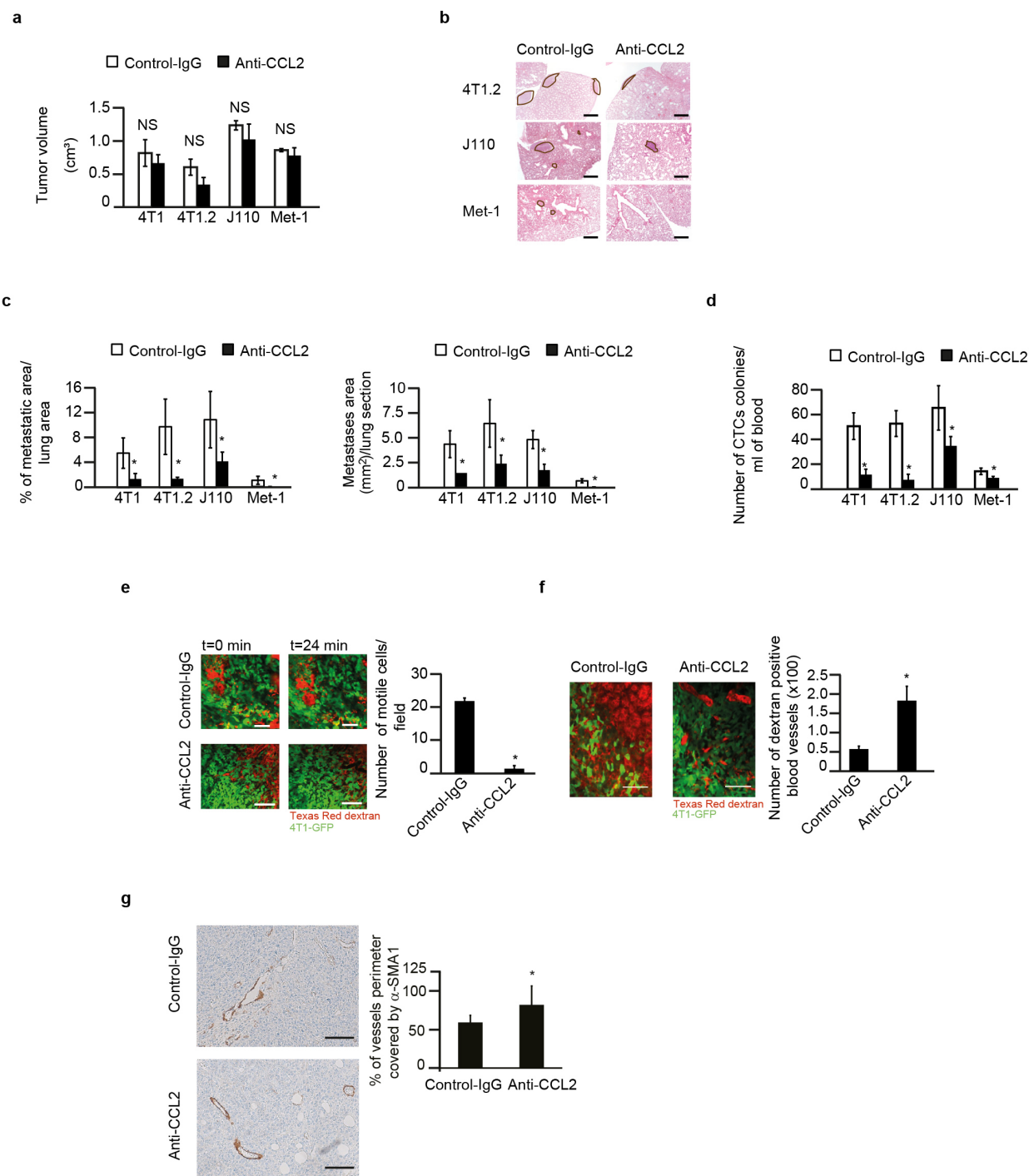
Immunohistochemistry on human samples was analysed with R/bioconductor<sup>22–25</sup> using a one-way ANOVA (assuming that the data are normally distributed). The different box plots represent the first to the third quartile of the data. The thick line is the median, and the whiskers in the box plot extend to the minimal and maximal values, after outliers have been removed. Outliers were defined by default as all data points that are more than 1.5-fold the box length (the interquartile range) away from the median.

22. The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
23. Zhu, J. *et al.* The UCSC Cancer Genomics Browser. *Nature Methods* **6**, 239–240 (2009).
24. Goldman, M. *et al.* The UCSC Cancer Genomics Browser: update 2013. *Nucleic Acids Res.* **41**, D949–D954 (2013).
25. Schmittgen, T. D. Real-time quantitative PCR. *Methods* **25**, 383–385 (2001).
26. Aslakson, C. J. & Miller, F. R. Selective events in the metastatic process defined by analysis of the sequential dissemination of subpopulations of a mouse mammary tumor. *Cancer Res.* **52**, 1399–1405 (1992).
27. Swirski, F. K. *et al.* Identification of splenic reservoir monocytes and their deployment to inflammatory sites. *Science* **325**, 612–616 (2009).
28. Wyckoff, J., Gligorijevic, B., Entenberg, D., Segall, J. & Condeelis, J. High-resolution multiphoton imaging of tumors *in vivo*. *Cold Spring Harb. Protoc.* **2011**, 1167–1184 (2011).
29. Abramoff, M. D., Magelhaes, P. J. & Ram, S. J. Image Processing with Image. *J. Biophotonics Intl* **11**, 36–42 (2004).
30. Egawa, G. *et al.* Intravital analysis of vascular permeability in mice using two-photon microscopy. *Sci. Rep.* **3**, 1932 (2013).



**Extended Data Figure 1 | CCL2 is overexpressed in breast cancer.** **a**, Positive correlation between immunohistochemical staining score of CCL2 and the number of CD68-positive macrophages per high-power field in human biopsies ( $n = 30$ ).  $P = 0.024$ ; data were analysed using R with one-way ANOVA (assuming that the data are normally distributed). The box plot represents the first to the third quartile of the data. Thick line indicates median, the whiskers extend to the minimal and maximal values. **b**, Kaplan–Meier survival curves showing overall survival of breast cancer patients with high or low CCL2 expression.  $P = 0.0372$ , log-rank test. **c**, CCL2 expression in

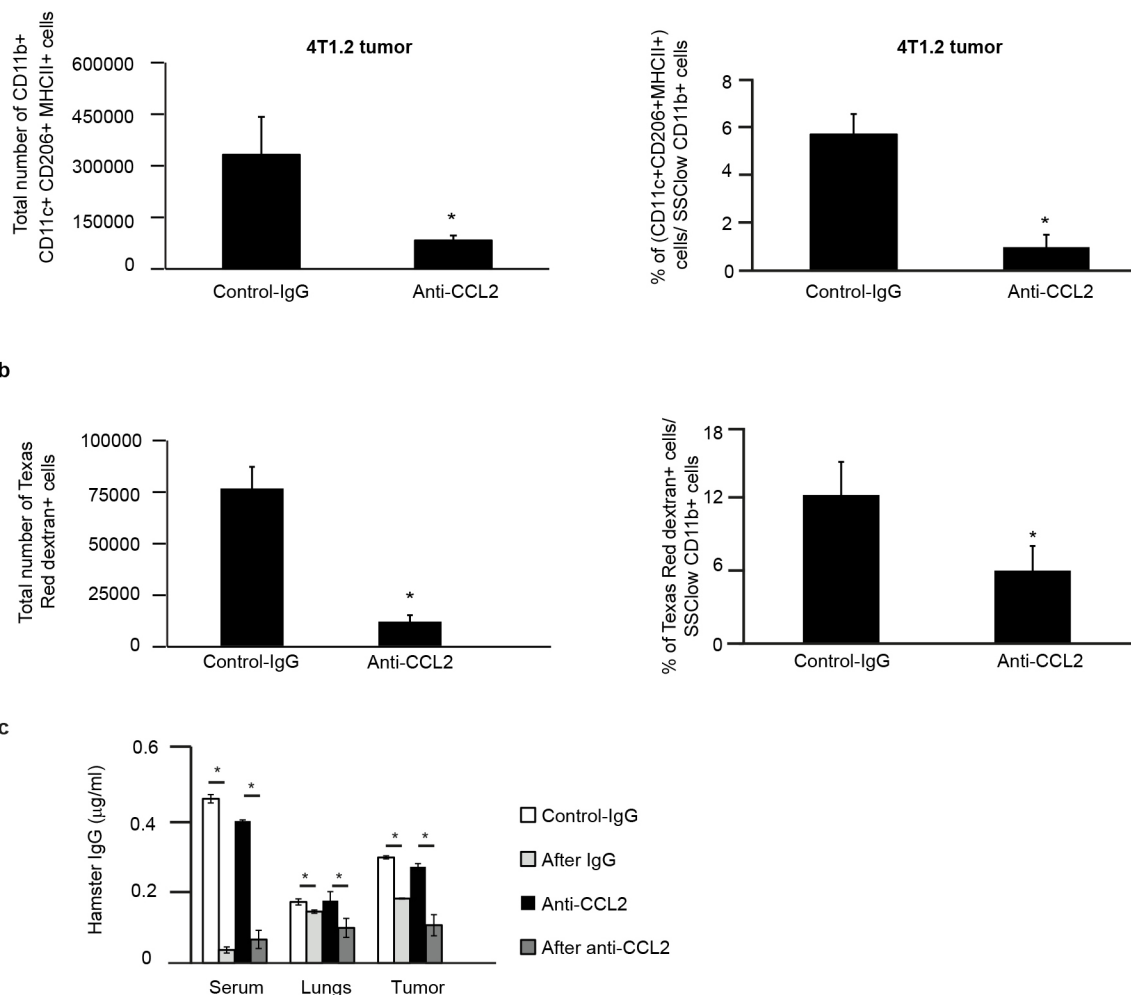
supernatants of murine breast cancer cell lines grown *in vitro*. Data are shown as means  $\pm$  s.e.m.,  $n = 3$  wells, one representative of three independent experiments. **d**, CCL2 expression in sera of mice bearing the indicated tumours for 14 days. Data are shown as means  $\pm$  s.e.m.,  $n = 3$  animals per group, one representative of two independent experiments. **e**, CCL2 levels in sera (left) and lung homogenate (right) from 4T1.2-tumour-bearing mice during anti-CCL2 or IgG treatment or 10 days after the cessation of treatment. Data are shown as means  $\pm$  s.e.m.,  $n = 5$  mice per group, pooled data from two independent experiments.  $*P < 0.05$ , unpaired *t*-test.



**Extended Data Figure 2 | Anti-CCL2 treatment reduces breast cancer metastases.** **a**, Tumour volumes on day 11 of treatment with anti-CCL2 or IgG. Data are shown as mean tumour volumes ± s.e.m.,  $n = 10$  mice, pooled data from two independent experiments. NS, not significant (unpaired  $t$ -test). **b**, Representative images of haematoxylin and eosin (H&E)-stained lung sections out of 20 images per group. Black circles indicate metastases. Scale bar, 500  $\mu$ m. **c**, Quantification of metastases as percentage of metastatic area per lung area (left) or as metastatic area per lung section (right). Data are shown as means ± s.e.m. of 20 fields of view on 5 sections per animal,  $n = 4$  mice per group, pooled data from two experiments. \* $P < 0.05$ , unpaired  $t$ -test. **d**, Effect of anti-CCL2 and IgG treatment on CTCs from tumour-bearing animals on day 11 of treatment. Data are shown as means ± s.e.m. of 20 fields of view on 5 sections per animal,  $n = 4$  mice per group, pooled data from two experiments. \* $P < 0.05$ , unpaired  $t$ -test. **e**, Intravital multiphoton images of 4T1-GFP primary tumours on day 7 of treatment with anti-CCL2 antibody or IgG. Left, representative two-dimensional images out of 24 fields from primary tumours

(green) 50  $\mu$ m below the tumour surface, 24 min after intravenous injection of 70 kDa Texas Red-dextran (red). Scale bar, 100  $\mu$ m. Right, mean numbers of motile tumour cells ± s.e.m. per 30 min acquisition, 6 fields per mouse,  $n = 4$  mice per group. \* $P < 0.05$ , unpaired  $t$ -test. **f**, Intravital multiphoton images of vasculature in 4T1-GFP primary tumours on day 7 of treatment. Left, representative two-dimensional images out of 12 fields per group, 50  $\mu$ m below the tumour surface, 15 min after intravenous injection of 70 kDa Texas Red-dextran (red). Scale bar, 100  $\mu$ m. Right, mean number of dextran-positive blood vessels counted over 42 min, starting 15 min after injection of Texas Red-dextran. Results are shown as means ± s.e.m.,  $n = 4$  mice per group and 3 fields per animal. \* $P < 0.05$ , unpaired  $t$ -test. **g**, Left, representative images out of 12 images per group from  $\alpha$ -SMA1-stained 4T1.2 tumour sections. Scale bar, 500  $\mu$ m. Right, quantification of the perimeter of blood vessels covered by  $\alpha$ -SMA1 staining. Data are shown as percentage vessel perimeter covered by  $\alpha$ -SMA1 ± s.e.m. of 12 fields of view on 4 sections per animal,  $n = 3$  mice per group, pooled data from two experiments. \* $P < 0.05$ , unpaired  $t$ -test.

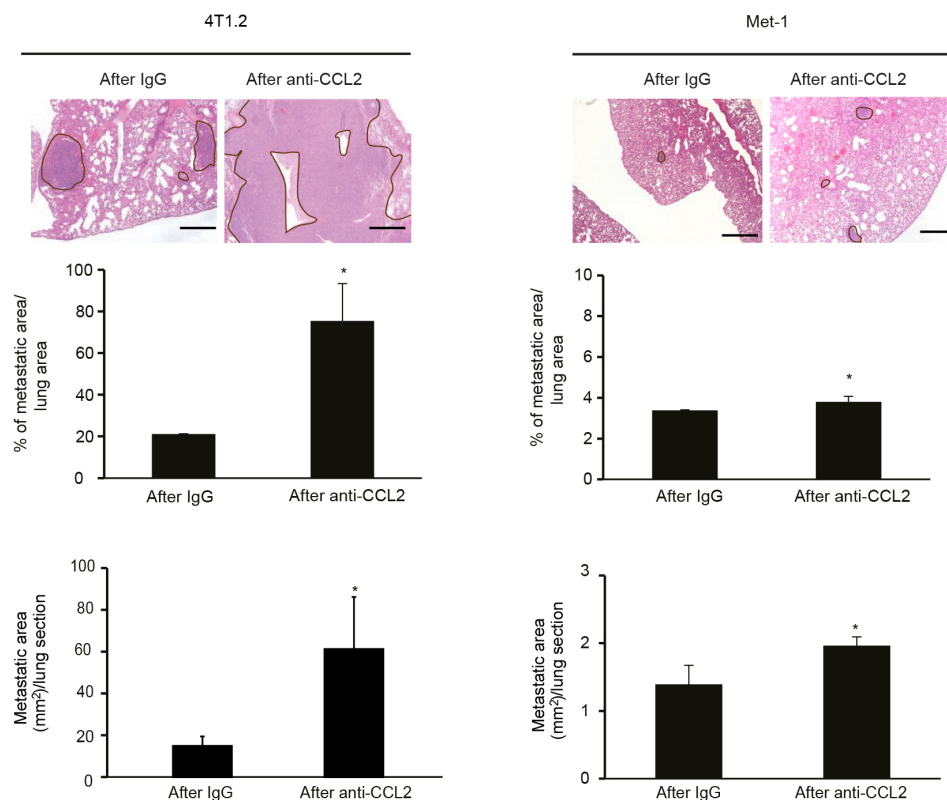




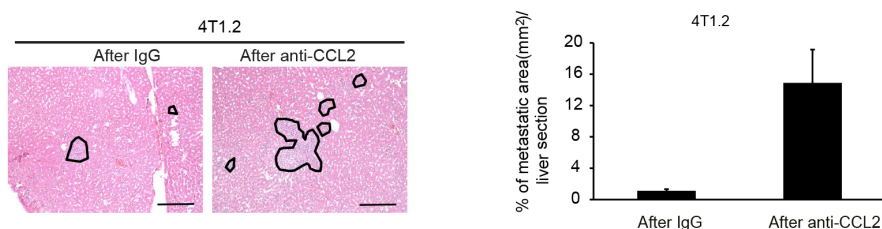
**Extended Data Figure 3 | Macrophages decline in primary 4T1.2 tumours upon treatment with anti-CCL2.** **a**, Left, total number of SSC<sup>low</sup>CD11b<sup>+</sup>CD11c<sup>+</sup>CD206<sup>+</sup>MHCII<sup>+</sup> macrophages. Right, percentage of CD11c<sup>+</sup>CD206<sup>+</sup>MHCII<sup>+</sup> macrophages per SSC<sup>low</sup>CD11b<sup>+</sup> cell population. Data are shown as means  $\pm$  s.e.m.,  $n = 3$  mice. \* $P < 0.05$ , unpaired  $t$ -test. **b**, Left, total number of Texas-Red-positive SSC<sup>low</sup>CD11b<sup>+</sup> macrophages. Right, percentage of Texas-Red-positive macrophages per SSC<sup>low</sup>CD11b<sup>+</sup> cells,

enumerated by flow cytometry of 4T1.2 primary tumours from tumour-bearing mice treated for 14 days, after intravenous injection of Texas Red-dextran. Data are shown as means  $\pm$  s.e.m.,  $n = 6$  mice pooled data of two independent experiments. \* $P < 0.05$ , unpaired  $t$ -test. **c**, Concentration of hamster IgG in serum, lung homogenates and tumour during treatment (day 14) and at 10 days after cessation of treatment as detected by ELISA. Data are shown as means  $\pm$  s.e.m.,  $n = 4$  mice per group. \* $P < 0.05$ , \*\* $P < 0.001$ , unpaired  $t$ -test.

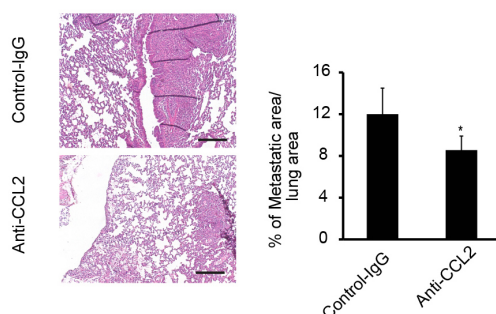
a



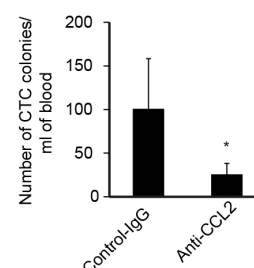
b



c



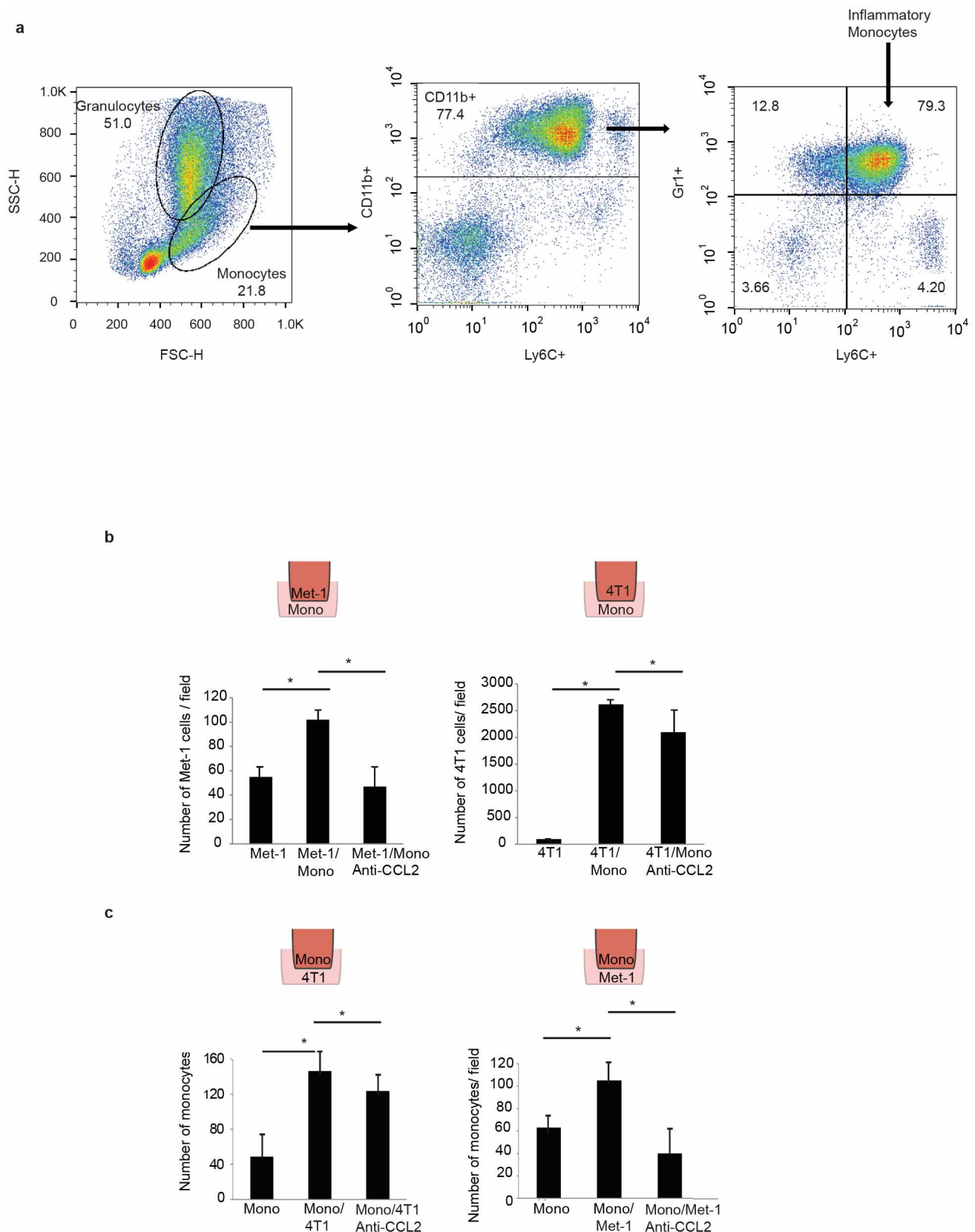
d



**Extended Data Figure 4 | Cessation of anti-CCL2 treatment increases lung metastases.** **a**, Top, representative H&E-stained lung sections from 4T1.2- and Met-1-tumour-bearing mice 10 days after cessation of treatment. Scale bar, 500  $\mu$ m. Bottom, quantification of lung metastases as percentage of metastatic area per lung area and metastatic area per lung section. Data are shown as means  $\pm$  s.e.m. of 20 fields of view on 5 lungs sections per animal,  $n = 4$  mice per group, pooled data from two experiments.  $*P < 0.05$ , unpaired  $t$ -test. One representative image out of 20 is shown per group. **b**, Left, representative H&E-stained liver sections from 4T1.2-tumour-bearing mice 10 days after cessation of treatment. Scale bar, 500  $\mu$ m. Right, quantification of liver metastases as metastatic area per lung section. Data are shown as means  $\pm$  s.e.m. of  $n = 10$

fields of view on 5 sections per animal,  $n = 2$  mice per group.  $*P < 0.05$ , unpaired  $t$ -test. One representative image out of 10 is shown per group.

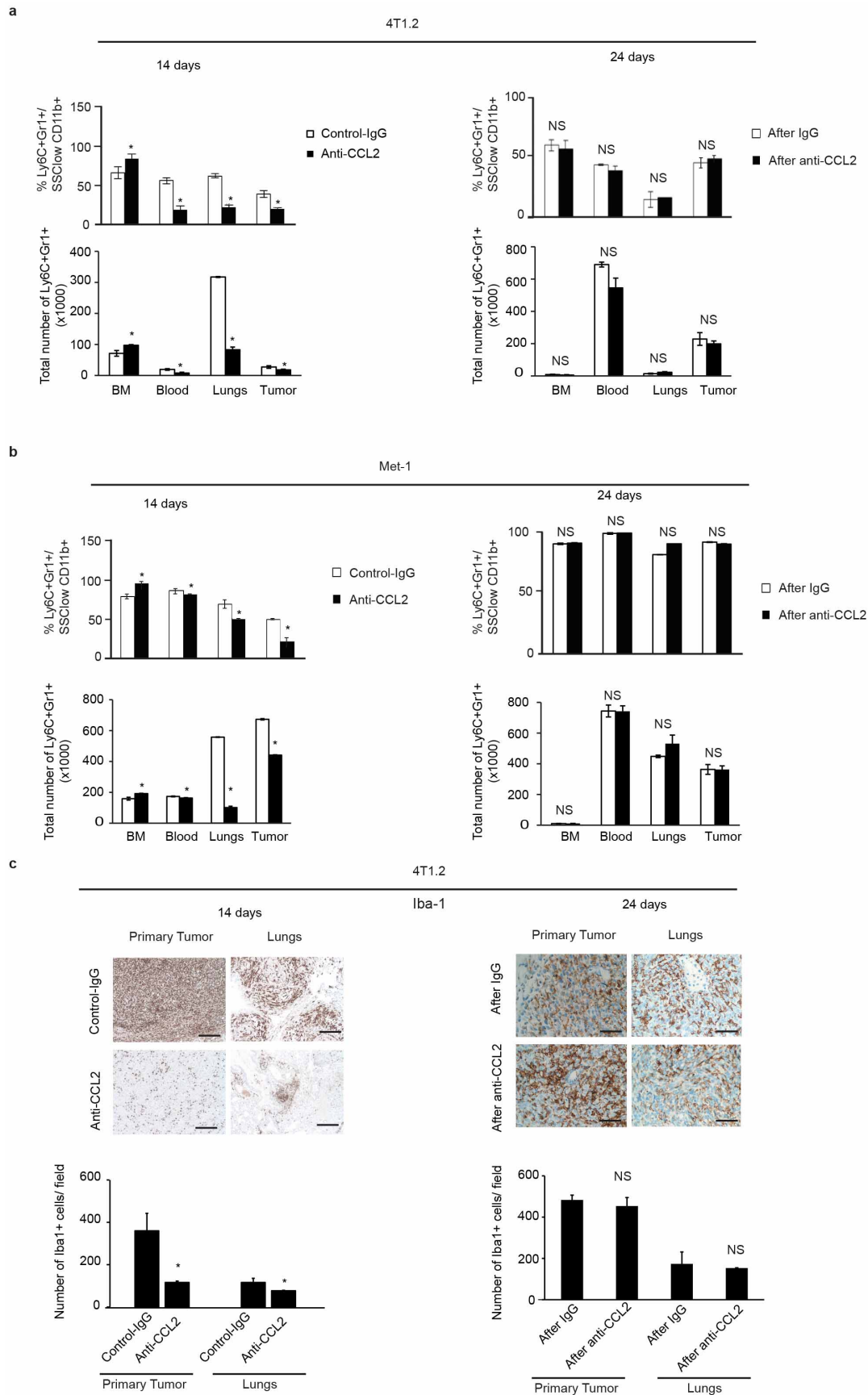
**c, d**, 4T1.2-tumour-bearing animals were treated for 24 consecutive days with anti-CCL2 or IgG control. **c**, Left, representative images of H&E-stained lung sections. Scale bar, 500  $\mu$ m. Right, lung metastases were quantified as percentage of metastatic area per lung area on day 24 of treatment. Data are shown as means  $\pm$  s.e.m. of 20 fields of view on 5 lungs sections per animal,  $n = 4$  mice per group.  $*P < 0.05$ , unpaired  $t$ -test. One representative image out of 20 is shown per group. **d**, Number of colonies formed by CTCs per ml blood collected on day 24 of treatment. Data are shown as means  $\pm$  s.e.m. of  $n = 4$  mice per group.  $*P < 0.05$ , unpaired  $t$ -test.



**Extended Data Figure 5 | CCL2 drives mutual attraction of monocytes and tumour cells.** **a**, Fluorescence-activated cell sorting (FACS) plots showing the gating strategy for identification and isolation of inflammatory monocytes. **b**, Transwell invasion assay of Met-1 cells (left) or 4T1 cells (right) on top of matrigel towards Ly6C<sup>+</sup> monocytes (Mono). Data are shown as means of invading cell numbers  $\pm$  s.e.m.,  $n = 4$  independent experiments with each

experiment performed in triplicate.  $*P < 0.05$ , ANOVA with Bonferroni post-test. **c**, Transwell invasion assay of Ly6C<sup>+</sup> monocytes on top of matrigel towards 4T1 (left) or Met-1 (right) cells. Data are shown as means of invading cell numbers  $\pm$  s.e.m.,  $n = 3$  independent experiments with each experiment performed in triplicate.  $*P < 0.05$ , ANOVA with post-hoc Bonferroni correction.

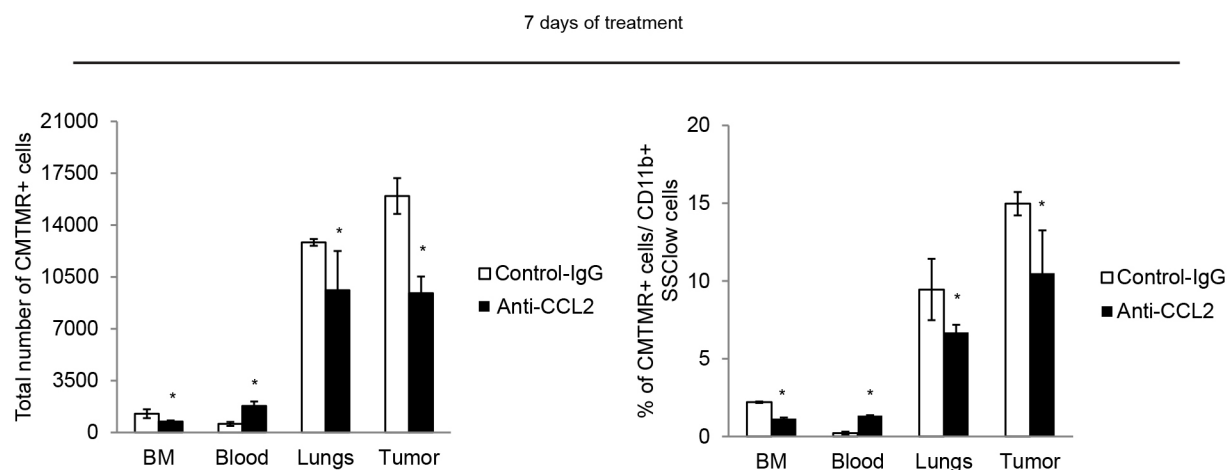




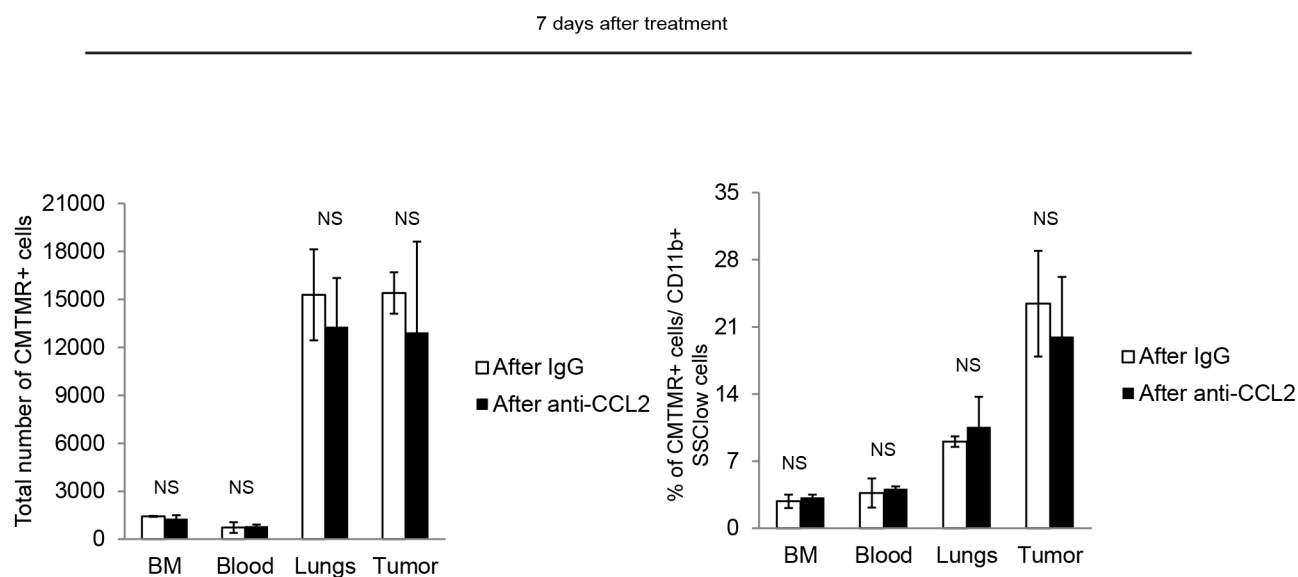
**Extended Data Figure 6 | Distribution of inflammatory monocytes assessed by flow cytometry during and after treatment *in vivo*.** **a**, Top, quantification of Ly6C<sup>+</sup>Gr1<sup>+</sup> monocytes as percentage of SSC<sup>low</sup>CD11b<sup>+</sup> cells on day 14 of anti-CCL2 or IgG treatment and 10 days after treatment interruption in 4T1.2-tumour-bearing animals. Data are shown as means  $\pm$  s.e.m.,  $n = 8$  mice, pooled data from two independent experiments.  $*P < 0.05$ , unpaired  $t$ -test. Bottom, quantification of total number of monocytes on day 14 of anti-CCL2 or IgG treatment and 10 days after treatment interruption in 4T1.2-tumour-bearing animals. Data are shown as means  $\pm$  s.e.m.,  $n = 8$  mice, pooled data from two independent experiments.  $*P < 0.05$ , unpaired  $t$ -test. **b**, Top, quantification of monocytes as Ly6C<sup>+</sup>Gr1<sup>+</sup> monocytes as percentage of SSC<sup>low</sup>CD11b<sup>+</sup> cells on day 14 of anti-CCL2 or IgG treatment and 10 days after treatment interruption in Met-1-tumour-bearing animals. Data are shown as means  $\pm$  s.e.m.,  $n = 8$  mice, pooled data from two independent experiments.  $*P < 0.05$ , unpaired

$t$ -test. Bottom, quantification of total number of monocytes on day 14 of anti-CCL2 or IgG treatment and 10 days after treatment interruption in 4T1.2-tumour-bearing animals. Data are shown as means  $\pm$  s.e.m.,  $n = 8$  mice, pooled data from two independent experiments.  $*P < 0.05$ , unpaired  $t$ -test. **c**, Top left, representative images of Iba1-stained lung and tumour sections on day 14 of treatment. Bottom left, quantification of Iba1 staining. Total numbers of Iba1-positive cells are shown  $\pm$  s.e.m. of 10 fields of view on 3 lungs sections per animal,  $n = 3$  mice per group. Top right, representative images of Iba1-stained lung and tumour sections 10 days after cessation of the treatment. Bottom right, quantification of Iba1 staining. Total numbers of Iba1-positive cells are shown  $\pm$  s.e.m. of 10 fields of view on 3 lung sections per animal,  $n = 3$  mice per group.  $*P < 0.05$ , ANOVA with post-hoc Bonferroni correction. Scale bar, 100  $\mu$ m. One representative image out of 10 is shown per group. NS, not significant.

a



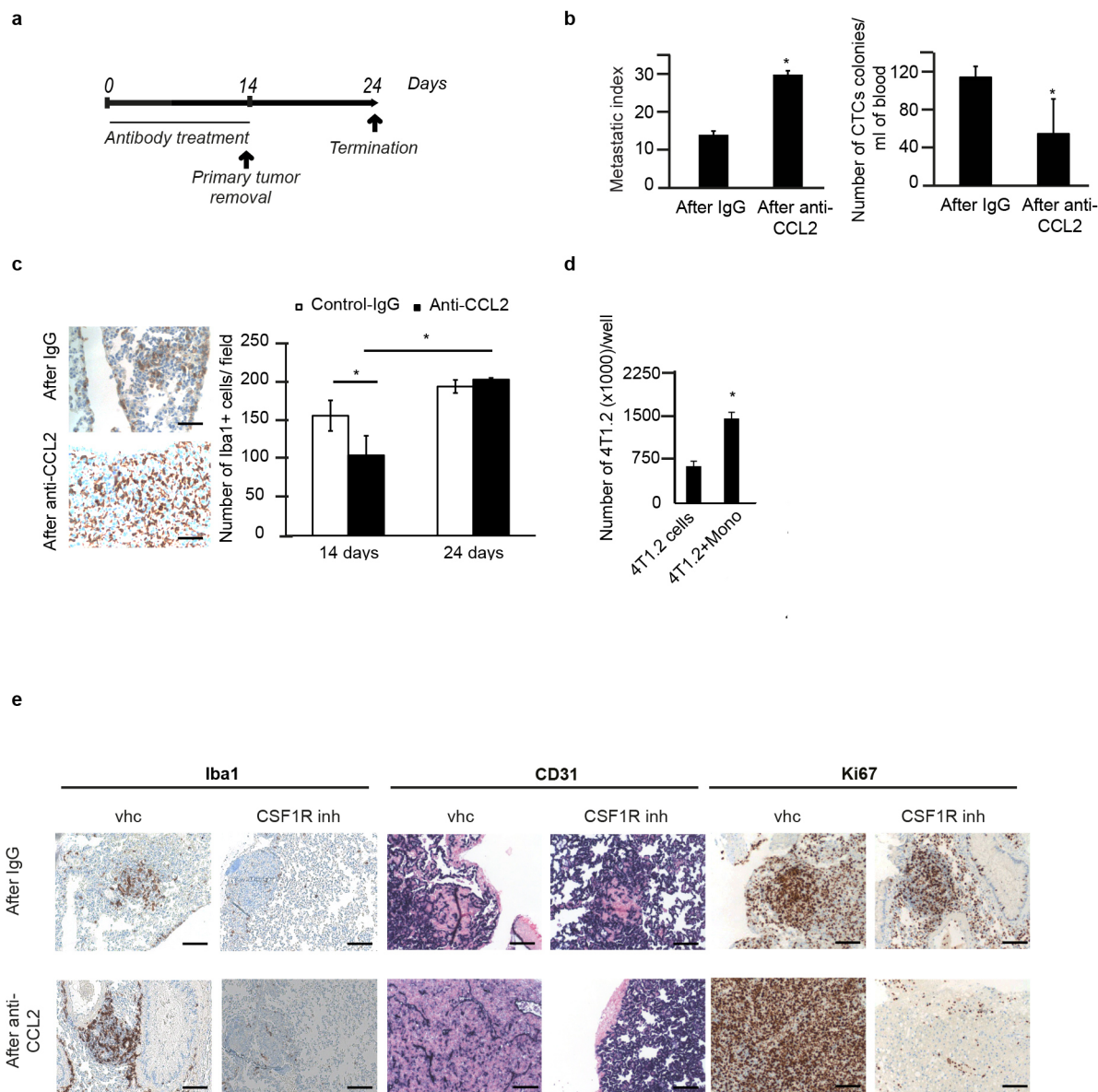
b



**Extended Data Figure 7 | Recruitment of adoptively transferred CMTMR-labelled CCR2<sup>+</sup> monocytes to the primary tumour and metastatic lung depends on CCL2.** **a**, Quantification of adoptively transferred CMTMR<sup>+</sup> monocytes as total numbers or percentages of SSC<sup>low</sup>CD11b<sup>+</sup> cells in different organs from tumour-bearing animals on day 7 of treatment. BM, bone marrow. Data are shown as means  $\pm$  s.e.m.,  $n = 5$  mice per group, one representative of

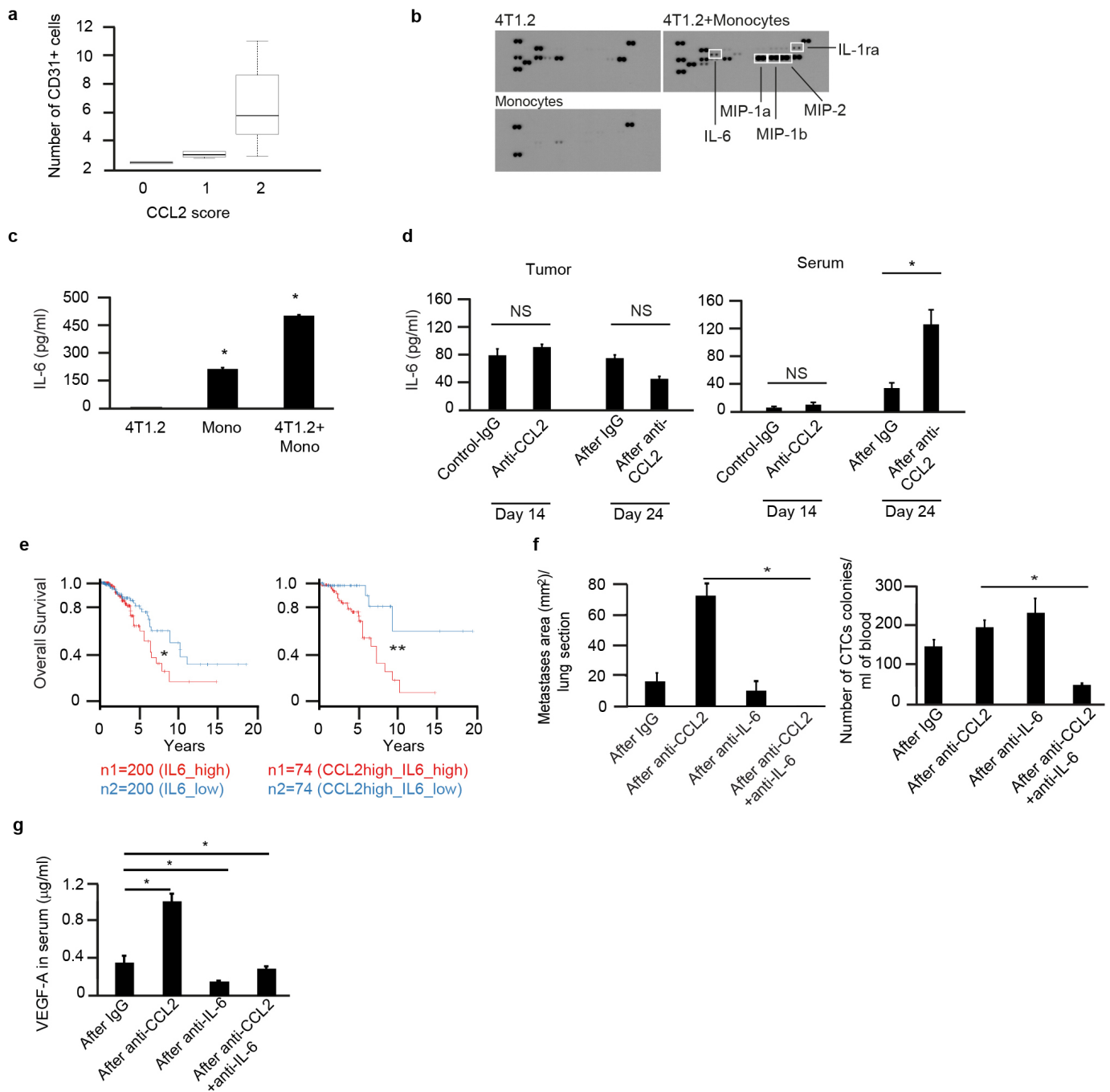
two independent experiments. \* $P < 0.05$ , unpaired  $t$ -test. **b**, Quantification of adoptively transferred CMTMR<sup>+</sup> monocytes as total numbers or percentages of SSC<sup>low</sup>CD11b<sup>+</sup> cells in different organs from tumour-bearing animals on day 7 after interruption of the treatments. Data are shown as means  $\pm$  s.e.m.,  $n = 3$  mice per group, one representative of two independent experiments. \* $P < 0.05$ , unpaired  $t$ -test. NS, not significant.





**Extended Data Figure 8 | Monocytes support growth of tumour cells *in vivo* and *in vitro*.** **a**, Timeline of the experiments. **b**, Left, metastatic index in lungs from 4T1.2-tumour-bearing animals on day 24. Data are shown as means  $\pm$  s.e.m.,  $n = 4$  mice per group, one representative of two independent experiments.  $*P < 0.05$ , unpaired *t*-test. Scale bar, 100  $\mu$ m. Right, the number of tumour cell colonies per ml blood of tumour-bearing animals on day 24 and after removal of the primary tumour. Data are shown as means  $\pm$  s.e.m.,  $n = 4$  mice per group, one representative of two independent experiments.  $*P < 0.05$ , unpaired *t*-test. **c**, Left, representative images of Iba1-stained lung sections in animals treated as in Extended Data Fig. 8a. Right, quantification of Iba1 staining on day 14 (last day of treatment and the day of tumour removal)

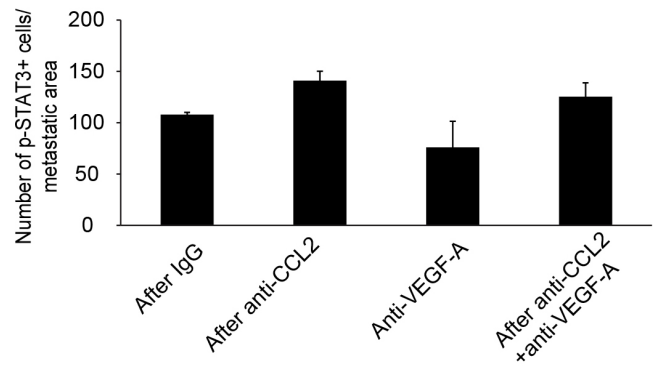
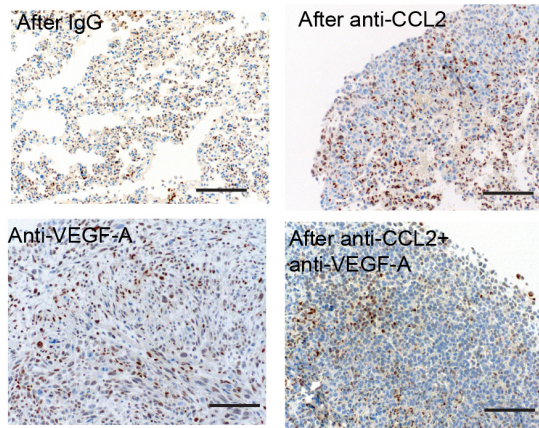
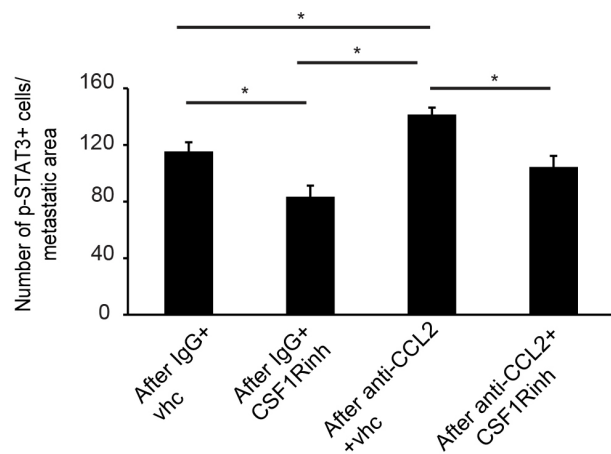
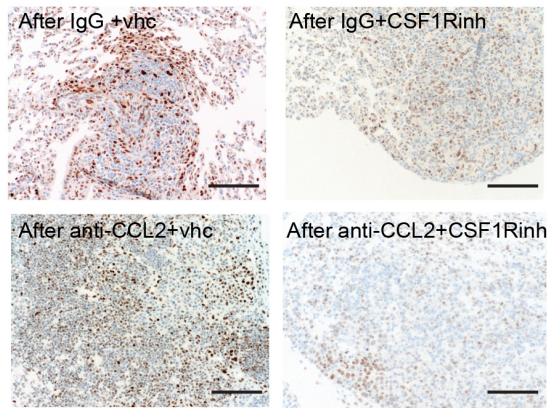
and on day 24 (10 days after stopping treatment and tumour removal). Total numbers of Iba1-positive cells are shown  $\pm$  s.e.m. of 10 fields of view on 3 lungs sections per animal,  $n = 3$  mice per group.  $*P < 0.05$ , unpaired *t*-test. Scale bar, 100  $\mu$ m. One representative image out of 10 is shown per group. **d**, Quantification of 4T1.2 cell viability using Trypan Blue 48 h after co-culture with sorted monocytes (Mono). Data are shown as means  $\pm$  s.e.m.,  $n = 3$ .  $*P < 0.05$ , unpaired *t*-test. **e**, Representative images of Iba1, CD31 and Ki67 staining of lung sections from animals treated with CSF1R inhibitor (CSF1Rinh) or vehicle (vhc) after treatment with anti-CCL2 or IgG control. One representative image out of 10 is shown per group.



### Extended Data Figure 9 | Upregulation of CCL2 and IL-6 reduces overall survival in breast cancer patients.

**a**, Correlation between immunohistochemical score of CCL2 and number of CD31<sup>+</sup> vessels per high-power field in biopsies from patients with breast carcinoma ( $n = 17$ ),  $P = 0.031$ . Data were analysed using R with one-way ANOVA (assuming that the data are normally distributed). The box plot represents the first to the third quartile of the data. Thick line indicates median, the whiskers in the box plot extend to the minimal and maximal values. **b**, Cytokine array showing expression of IL-6 in the supernatant of mono- or co-cultures of 4T1.2 cells and primary monocytes, in the presence or absence of anti-CCL2. **c**, Quantification of IL-6 expression in cell supernatants of monocytes in mono- or co-culture with 4T1.2 by ELISA. Data are shown as means  $\pm$  s.e.m.,  $n = 4$  independent experiments each with biological triplicates.  $*P < 0.05$ , ANOVA with post-hoc Bonferroni correction. **d**, IL-6 expression levels in primary tumours (left) or in serum (right) on day 14 of treatment and after cessation of treatment (day 24). Data are shown as means  $\pm$  s.e.m.,  $n = 6$  mice per group, pooled data from two

independent experiments.  $*P < 0.05$ , ANOVA with post-hoc Bonferroni correction. **e**, Left, Kaplan-Meier survival curves showing overall survival of 400 breast tumour patients expressing either low (blue) or high (red) IL-6 levels.  $*P < 0.045$ , log-rank test. Right, Kaplan-Meier survival curves showing overall survival of 148 patients with tumours expressing either CCL2 high/IL-6 low (blue) or CCL2 high/IL-6 high (red) levels.  $*P < 0.0017$ , log-rank test. **f**, Left, quantification of lung metastases. Metastatic area per lung section is shown  $\pm$  s.e.m. of 20 fields of view on 5 lungs sections per animal,  $n = 4$  mice per group.  $*P < 0.05$ , ANOVA with post-hoc Bonferroni correction. Right, number of colonies formed by CTCs per ml blood in animals treated as in Fig. 3c on day 24. Data are shown as means  $\pm$  s.e.m.,  $n = 10$  mice per group, pooled data from two experiments.  $*P < 0.05$ , ANOVA with post-hoc Bonferroni correction. **g**, VEGF-A expression as measured by ELISA in serum from 4T1.2-tumour-bearing animals treated as in Fig. 3c. Data are shown as means  $\pm$  s.e.m.,  $n = 4$  mice per group, pooled data from two experiments.  $*P < 0.05$ , ANOVA with post-hoc Bonferroni correction. NS, not significant.

**a****b**

**Extended Data Figure 10 | Monocytes/macrophages induce p-STAT3 in metastatic lungs upstream of VEGF-A expression.** **a**, Left, representative images of p-STAT3-stained lung sections from animals treated as in Fig. 4a. Scale bar, 100  $\mu$ m. Right, quantification of p-STAT3 staining. Total numbers of p-STAT3-positive cells per field are shown  $\pm$  s.e.m. of 10 fields of view on 3 lungs sections per animal,  $n = 3$  mice per group. **b**, Left, representative images

of p-STAT3-stained lung sections from animals treated as in Fig. 2c. Scale bar, 100  $\mu$ m. Right, quantification of p-STAT3 staining. Total numbers of p-STAT3-positive cells per field are shown  $\pm$  s.e.m. of 10 fields of view on 3 lungs sections per animal,  $n = 3$  mice per group. CSF1Rinh, CSF1R inhibitor; vhc, vehicle.  $*P < 0.05$ , ANOVA with post-hoc Bonferroni correction. **a**, **b**, One representative image out of 10 is shown per group.

# Tumour-infiltrating Gr-1<sup>+</sup> myeloid cells antagonize senescence in cancer

Diletta Di Mitri<sup>1\*</sup>, Alberto Toso<sup>1\*</sup>, Jing Jing Chen<sup>1,2</sup>, Manuela Sarti<sup>1</sup>, Sandra Pinton<sup>1</sup>, Tanja Rezzonico Jost<sup>3</sup>, Rocco D'Antuono<sup>3</sup>, Erica Montani<sup>3</sup>, Ramon Garcia-Escudero<sup>1,4</sup>, Ilaria Guccini<sup>1</sup>, Sabela Da Silva-Alvarez<sup>5</sup>, Manuel Collado<sup>5</sup>, Mario Eisenberger<sup>6</sup>, Zhe Zhang<sup>7</sup>, Carlo Catapano<sup>1</sup>, Fabio Grassi<sup>3,8</sup> & Andrea Alimonti<sup>1,2</sup>

**Aberrant activation of oncogenes or loss of tumour suppressor genes opposes malignant transformation by triggering a stable arrest in cell growth, which is termed cellular senescence<sup>1–3</sup>. This process is finely tuned by both cell-autonomous and non-cell-autonomous mechanisms that regulate the entry of tumour cells to senescence<sup>4–6</sup>. Whether tumour-infiltrating immune cells can oppose senescence is unknown. Here we show that at the onset of senescence, PTEN null prostate tumours in mice<sup>2,7</sup> are massively infiltrated by a population of CD11b<sup>+</sup>Gr-1<sup>+</sup> myeloid cells that protect a fraction of proliferating tumour cells from senescence, thus sustaining tumour growth. Mechanistically, we found that Gr-1<sup>+</sup> cells antagonize senescence in a paracrine manner by interfering with the senescence-associated secretory phenotype of the tumour through the secretion of interleukin-1 receptor antagonist (IL-1RA). Strikingly, *Pten*-loss-induced cellular senescence was enhanced *in vivo* when *Il1ra* knockout myeloid cells were adoptively transferred to PTEN null mice. Therapeutically, docetaxel-induced senescence and efficacy were higher in PTEN null tumours when the percentage of tumour-infiltrating CD11b<sup>+</sup>Gr-1<sup>+</sup> myeloid cells was reduced using an antagonist of CXC chemokine receptor 2 (CXCR2)<sup>8</sup>. Taken together, our findings identify a novel non-cell-autonomous network, established by innate immunity, that controls senescence evasion and chemoresistance. Targeting this network provides novel opportunities for cancer therapy.**

Cellular senescence is a stable state of cell growth arrest that opposes tumour initiation and progression in a variety of *in vivo* tumour models<sup>1,2</sup>. Recent studies have revealed an unexpected role for both adaptive and innate immunity in the regulation of cellular senescence. Immune cells can either clear senescent cells from tumours or induce senescence in cancer cells by secreting pro-inflammatory cytokines<sup>9–11</sup>. However, whether tumour-infiltrating immune cells can also oppose senescence *in vivo* is not known. We have previously shown that complete inactivation of the tumour suppressor gene *Pten* in the mouse prostate epithelium induces the formation of benign tumours characterized by a strong senescence response that opposes tumour progression<sup>2,7</sup>. However, these tumours grow over time and progress to become more aggressive and invasive tumours<sup>2,12</sup>. Indeed, at the onset of senescence (7–8 weeks), PTEN null tumours (hereafter referred to as *Pten*<sup>pc-/-</sup> tumours) are characterized by the concomitant presence of both senescent and proliferative cellular compartments, as shown by the expression of p16<sup>INK4A</sup> (also known as CDKN2A), pHP1γ (also known as CBX3) and senescence-associated β-galactosidase (SA-β-gal) and by Ki-67 positivity, respectively (Fig. 1a and Extended Data Fig. 1a–c). This finding suggests that from the early stages of tumorigenesis, a fraction of proliferating tumour cells evades senescence. Given the interplay between senescent tumour cells and immune cells, we speculated that *Pten*<sup>pc-/-</sup> tumours may evade senescence

in a non-cell-autonomous manner and that the tumour microenvironment could be the source of factors that hinder the senescence response.

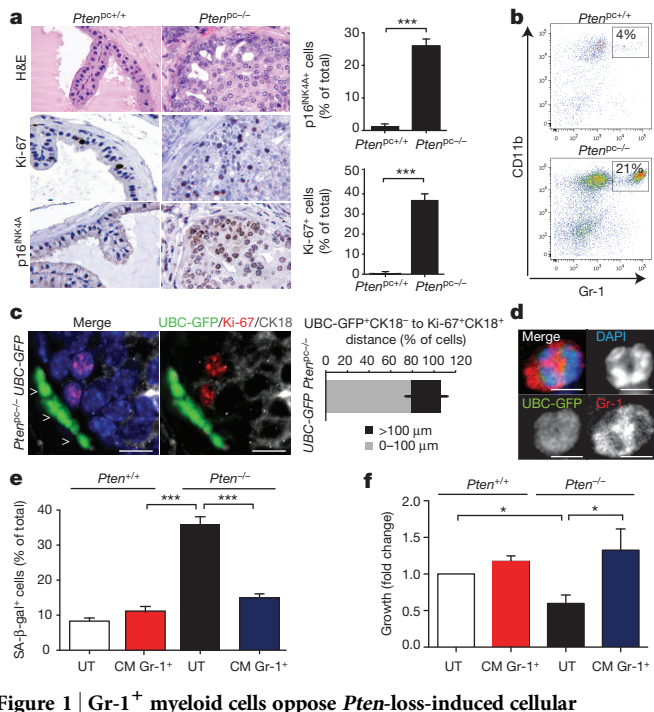
To address this hypothesis, we first characterized the immune microenvironment of *Pten*<sup>pc-/-</sup> tumours at the onset of senescence and found a strong infiltration of CD45<sup>+</sup>CD11b<sup>+</sup>Gr-1<sup>+</sup> myeloid cells (hereafter referred to as Gr-1<sup>+</sup> cells) (Fig. 1b and Extended Data Fig. 1d–f). Moreover, the prostate lobes with the highest percentage of Ki-67 staining were the most infiltrated by Gr-1<sup>+</sup> cells (Extended Data Fig. 1g). To study the localization of tumour-infiltrating Gr-1<sup>+</sup> cells in the prostate, we adoptively transferred bone marrow precursors from mice transgenic for green fluorescent protein (GFP) under control of the human ubiquitin C (*UBC*) promoter<sup>13</sup> into lethally irradiated *Pten*<sup>pc-/-</sup> mice (Extended Data Fig. 1h). Immunofluorescence showed that GFP<sup>+</sup> cells localized to the stroma and the prostate glands (Fig. 1c and Extended Data Fig. 1i, j). Moreover, GFP<sup>+</sup> cells were localized in close proximity to proliferating (Ki-67<sup>+</sup>) tumour cells (Fig. 1c and Extended Data Fig. 1j). Immune cells secrete a variety of cytokines that regulate senescence<sup>11</sup>. Interestingly, the majority of GFP<sup>+</sup> cells were spatially distributed within 100 μm of Ki-67<sup>+</sup> epithelial cells (Fig. 1c), suggesting that GFP<sup>+</sup> epithelial cells may interfere with senescence in a paracrine manner<sup>14,15</sup>. Notably, ~70% of the tumour-infiltrating GFP<sup>+</sup> cells expressed the myeloid differentiation antigen Gr-1 (Fig. 1d and Extended Data Fig. 1k). We also confirmed these results in tumour sections from non-irradiated *Pten*<sup>pc-/-</sup> mice<sup>16</sup> (Extended Data Fig. 1l). To assess whether factors secreted by Gr-1<sup>+</sup> cells can oppose *Pten*-loss-induced cellular senescence, we cultured *Pten*<sup>-/-</sup> mouse embryonic fibroblasts (MEFs), which undergo senescence *in vitro*<sup>2,7</sup>, in the presence of conditioned medium obtained from Gr-1<sup>+</sup> myeloid cells<sup>17</sup> (Extended Data Fig. 1m, n). Surprisingly, *Pten*-loss-induced cellular senescence was impaired in *Pten*<sup>-/-</sup> MEFs cultured in the presence of conditioned medium from Gr-1<sup>+</sup> cells (Fig. 1e and Extended Data Fig. 1o). These data demonstrate that Gr-1<sup>+</sup> cells oppose senescence in a paracrine manner.

To identify secreted factors that mediate the anti-senescence function of Gr-1<sup>+</sup> cells, we compared the cytokine profile of *Pten*<sup>pc-/-</sup> tumours before and after depletion of immune cells<sup>18</sup> and found that IL-1RA was the cytokine present at the most reduced level after immunodepletion (Extended Data Fig. 2a and Supplementary Table 1). Gene expression analysis, quantitative PCR with reverse transcription (qRT-PCR) and immunofluorescence confirmed that Gr-1<sup>+</sup> cells were the major source of IL-1RA in *Pten*<sup>pc-/-</sup> tumours relative to epithelia (Fig. 2a, b and Extended Data Fig. 2b). It should also be noted that CD11b<sup>+</sup>Gr-1<sup>+</sup>F4/80<sup>+</sup> cells released IL-1RA in the tumour microenvironment (Extended Data Fig. 2c). IL-1RA is an antagonist of IL-1R and has been reported to impair oncogene-induced senescence *in vitro* by blocking IL-1α-mediated signalling<sup>5,6</sup>. Notably, in *Pten*<sup>pc-/-</sup> tumours, the prostate lobes with the

<sup>1</sup>Institute of Oncology Research (IOR), Oncology Institute of Southern Switzerland, Bellinzona CH6500, Switzerland. <sup>2</sup>Faculty of Biology and Medicine, University of Lausanne UNIL, Lausanne CH1011, Switzerland. <sup>3</sup>Institute for Research in Biomedicine (IRB), Bellinzona CH6500, Switzerland. <sup>4</sup>Molecular Oncology Unit, CIEMAT, 28040 Madrid, Spain. <sup>5</sup>Laboratory of Stem Cells in Cancer and Aging, (stemCHUS) Health Research Institute of Santiago de Compostela (IDIS), Clinical University Hospital (CHUS), E15706 Santiago de Compostela, Spain. <sup>6</sup>Department of Oncology, Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University, Baltimore, Maryland 21231-1000, USA. <sup>7</sup>Divisions of BioStatistics, Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University, Baltimore, Maryland 21231-1000, USA. <sup>8</sup>Department of Medical Biotechnology and Translational Medicine, University of Milan, Milan I-20100, Italy.

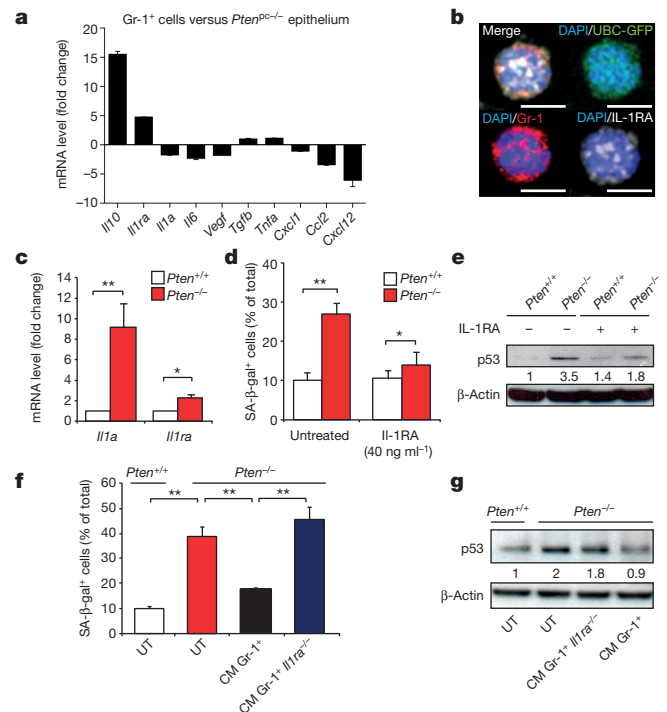
\*These authors contributed equally to this work.





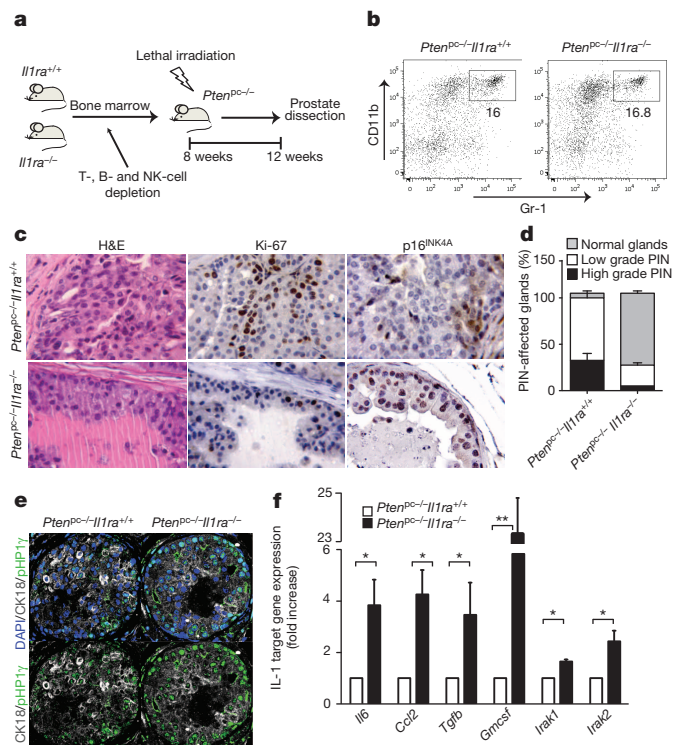
**Figure 1 | Gr-1<sup>+</sup> myeloid cells oppose *Pten*-loss-induced cellular senescence.** **a**, Haematoxylin and eosin (H&E), Ki-67 and p16<sup>INK4A</sup> immunohistochemical staining (Ki-67 and p16<sup>INK4A</sup>, blue; nuclei, brown) (left) and quantification (right) in tumours from 8-week-old *Pten*<sup>pc+/+</sup> and *Pten*<sup>pc-/-</sup> mice ( $n = 15$  mice, 1 tumour assessed per mouse, 3 sections assessed per tumour,  $\geq 5$  fields assessed per section). Original magnification,  $\times 400$ . **b**, Flow cytometry plots of CD11b<sup>+</sup>Gr-1<sup>+</sup> immune cells in tumours from 8-week-old *Pten*<sup>pc+/+</sup> and *Pten*<sup>pc-/-</sup> mice ( $n = 6$ ), gating on CD45<sup>+</sup> cells. **c**, Confocal images (left) and quantification (right) of the localization and distance between tumour-infiltrating myeloid cells (GFP<sup>+</sup>, green) and proliferating epithelial cells (cytokeratin 18 (CK18), grey; Ki-67, red) in *Pten*<sup>pc-/-</sup> UBC-GFP prostate lesions (nuclei, blue (DAPI)) ( $n = 4$  mice, 1 tumour per mouse, 5 fields acquired, 320 cells measured). The arrow heads point to positions where UBC-GFP<sup>+</sup> cells colocalize in close proximity to Ki-67<sup>+</sup> cells. Scale bar, 10  $\mu$ m. **d**, Representative confocal images of UBC-GFP<sup>+</sup>Gr-1<sup>+</sup> cells in *Pten*<sup>pc-/-</sup> UBC-GFP prostate lesions (nuclei, blue (DAPI)). Scale bar, 5  $\mu$ m. **e**, Quantification of SA- $\beta$ -gal staining in *Pten*<sup>+/+</sup> and *Pten*<sup>-/-</sup> cells ( $n = 5$ ). CM, conditioned medium; UT, untreated. **f**, Cell proliferation of *Pten*<sup>+/+</sup> and *Pten*<sup>-/-</sup> cells (fold change compared with UT *Pten*<sup>pc+/+</sup>) ( $n = 5$ ). **a**, **c**, **e**, **f**, Error bars, mean  $\pm$  s.e.m.  $P$  values were derived from an unpaired, two-tailed Student's  $t$ -test (\* $P < 0.05$ ; \*\*\* $P < 0.001$ ).

highest percentage of Gr-1<sup>+</sup> myeloid cells (Extended Data Fig. 1g) displayed the highest levels of IL-1RA and the lowest levels of p16<sup>INK4A</sup> (Extended Data Fig. 3a). To elucidate the role of IL-1 $\alpha$ -mediated signalling in *Pten*-loss-induced cellular senescence, we cultured *Pten*<sup>-/-</sup> MEFs in the presence of IL-1RA. While IL-1 $\alpha$  was strongly expressed in senescent *Pten*<sup>-/-</sup> cells, unlike in *Pten*<sup>+/+</sup> cells, the expression of IL-1RA was slightly higher than in *Pten*<sup>+/+</sup> cells (Fig. 2c). Remarkably, treatment with IL-1RA decreased both SA- $\beta$ -gal staining and levels of the tumour suppressor protein p53 in *Pten*<sup>-/-</sup> cells (Fig. 2d, e). *In-vitro*-polarized Gr-1<sup>+</sup> myeloid cells expressed high levels of *Il1ra* (Extended Data Fig. 3b). When *Pten*<sup>-/-</sup> MEFs were cultured in the presence of conditioned medium from Gr-1<sup>+</sup> myeloid cells derived from *Il1ra* knockout (*Il1ra*<sup>-/-</sup>) mice<sup>19,20</sup>, *Pten*-loss-induced cellular senescence was not impaired (Fig. 2f, g). Similar results were obtained in MEFs transfected with Ha-ras<sup>V12</sup> (H-ras), suggesting that myeloid cells also oppose oncogene-induced senescence (Extended Data Fig. 3c, d). Moreover, conditioned medium from Gr-1<sup>+</sup> cells pre-treated with a JAK2 inhibitor failed to block *Pten*-loss-induced cellular senescence<sup>21</sup> (Extended Data Fig. 3e, f). Importantly, IL-1RA also blocked docetaxel-induced senescence<sup>22</sup> in human prostate cancer cells (Extended Data Fig. 4a, b).



**Figure 2 | Gr-1<sup>+</sup> myeloid cells oppose *Pten*-loss-induced cellular senescence by interfering with IL-1 $\alpha$  signalling *in vitro*.** **a**, mRNA levels of secreted factors ( $n = 3$  per group). **b**, Representative confocal images of GFP<sup>+</sup> myeloid cells co-expressing Gr-1 (red) and IL-1RA (grey) (nuclei, blue (DAPI)). Scale bar, 5  $\mu$ m. **c**, *Il1a* and *Il1ra* mRNA levels in *Pten*<sup>+/+</sup> and *Pten*<sup>-/-</sup> MEFs ( $n = 3$ ). **d**, Percentage of SA- $\beta$ -gal<sup>+</sup> cells in *Pten*<sup>+/+</sup> and *Pten*<sup>-/-</sup> MEFs treated with recombinant IL-1RA. Numbers indicate fold changes in protein level. **e**, Western blot for p53 in MEFs treated with recombinant IL-1RA. Numbers indicate fold changes in protein level. **f**, Percentage of SA- $\beta$ -gal<sup>+</sup> cells in *Pten*<sup>+/+</sup> and *Pten*<sup>-/-</sup> MEFs cultured in the presence of conditioned medium from *Il1ra*<sup>+/+</sup>Gr-1<sup>+</sup> or *Il1ra*<sup>-/-</sup>Gr-1<sup>+</sup> cells ( $n = 5$ ). **g**, Western blot for p53. Numbers indicate fold changes in protein level. **a**, **c**, **d**, **f**, Error bars, mean  $\pm$  s.e.m.  $P$  values were derived from an unpaired, two-tailed Student's  $t$ -test (\* $P < 0.05$ ; \*\* $P < 0.01$ ).

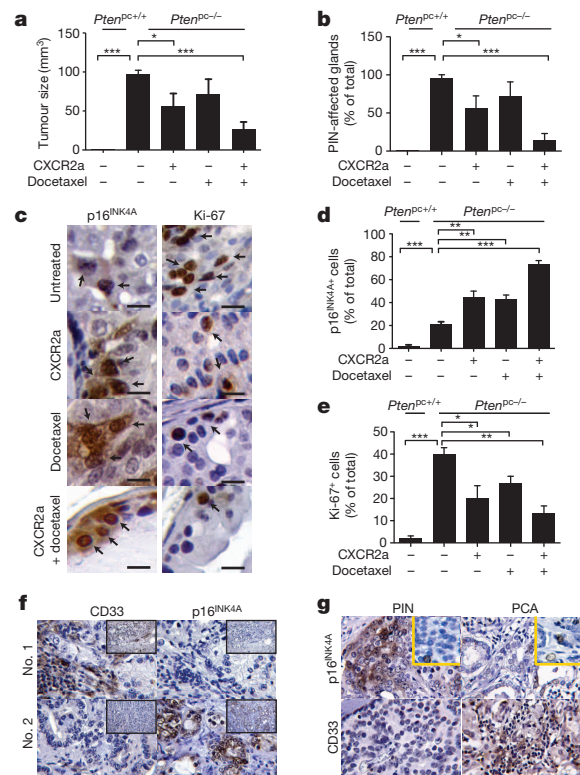
We next validated our findings in a mouse model of oncogene-induced senescence. *Ki-ras*<sup>G12V</sup> mice develop both lung adenomas and adenocarcinomas that display stage-dependent expression of senescence markers<sup>23</sup>. Interestingly, lung adenocarcinomas were characterized by the infiltration of Gr-1<sup>+</sup> myeloid cells, IL-1RA expression and Ki-67 positivity and the absence of senescence markers (Extended Data Fig. 5a). By contrast, senescent lung adenomas were poorly infiltrated by myeloid cells (Extended Data Fig. 5a–c). These data suggest that myeloid cells may oppose senescence in different types of tumour, irrespective of the genetic background. To validate our findings *in vivo*, we adoptively transferred bone marrow precursors from *Il1ra*<sup>+/+</sup> or *Il1ra*<sup>-/-</sup> mice into lethally irradiated *Pten*<sup>pc-/-</sup> mice (yielding *Pten*<sup>pc-/-</sup>*Il1ra*<sup>+/+</sup> mice and *Pten*<sup>pc-/-</sup>*Il1ra*<sup>-/-</sup> mice) (Fig. 3a and Extended Data Fig. 6a). Notably, *Pten*<sup>pc-/-</sup>*Il1ra*<sup>+/+</sup> and *Pten*<sup>pc-/-</sup>*Il1ra*<sup>-/-</sup> tumours were infiltrated equally by Gr-1<sup>+</sup> cells (Fig. 3b). Strikingly, histopathological analysis revealed that *Pten*<sup>pc-/-</sup>*Il1ra*<sup>-/-</sup> tumours displayed an almost complete normalization of glands affected by prostatic intraepithelial neoplasia (Fig. 3c, d), which was associated with decreased cell proliferation, increased senescence and absence of apoptosis (Fig. 3c, e and Extended Data Fig. 6b–h), in contrast to *Pten*<sup>pc-/-</sup>*Il1ra*<sup>+/+</sup> tumours. Finally, the enhanced senescence response in *Pten*<sup>pc-/-</sup>*Il1ra*<sup>-/-</sup> tumours was associated with the activation of IL-1 $\alpha$ -mediated signalling<sup>5</sup> (Fig. 3f). Myeloid cells were adoptively transferred to *Pten*<sup>pc-/-</sup> mice in the absence of T cells (Extended Data Fig. 7a, b), suggesting that Gr-1<sup>+</sup> myeloid cells oppose senescence through a novel pro-tumorigenic mechanism that does not involve this cell population<sup>8,24</sup>. These results demonstrate that Gr-1<sup>+</sup> myeloid cells antagonize senescence *in vivo* in a paracrine manner by interfering with



**Figure 3 | Adoptively transferred *Il1ra*<sup>-/-</sup> bone marrow precursors enhanced *Pten*-loss-induced cellular senescence *in vivo*.** **a**, Experimental set-up. **b**, Flow cytometry plots showing equal infiltration of Gr-1<sup>+</sup> cells in *Pten*<sup>PC-/-</sup>*Il1ra*<sup>+/+</sup> and *Pten*<sup>PC-/-</sup>*Il1ra*<sup>-/-</sup> mice. **c**, H&E, Ki-67 and p16<sup>INK4A</sup> immunohistochemical staining of sections from 12-week-old *Pten*<sup>PC-/-</sup>*Il1ra*<sup>+/+</sup> and *Pten*<sup>PC-/-</sup>*Il1ra*<sup>-/-</sup> mice. **d**, Quantification of prostatic intraepithelial neoplasia (PIN)-affected glands in *Pten*<sup>PC-/-</sup>*Il1ra*<sup>+/+</sup> and *Pten*<sup>PC-/-</sup>*Il1ra*<sup>-/-</sup> mice. **e**, Confocal images of staining for CK18 (grey) and pHP1γ (green) in prostate tumours from the indicated genotypes (nuclei, blue (DAPI)). Percentage of cells that stained positive for pHP1γ, 15 ± 7% (bottom left) and 37 ± 13% (bottom right). Original magnification, ×400. **f**, IL-1α signalling. Fold change in the expression of IL-1 target genes in tumours of the indicated genotypes. **d**, **f**, *n* = 4; 1 tumour per mouse; 3 sections per mouse; ≥5 fields per section. Error bars, mean ± s.e.m. *P* values were derived from an unpaired, two-tailed Student's *t*-test (\**P* < 0.05; \*\**P* < 0.01).

IL-1α-mediated signalling. Moreover, *Pten*<sup>PC-/-</sup> tumours from mice treated with IL-1α showed a significant reduction in the percentage of glands affected by prostatic intraepithelial neoplasia and the number of Ki-67<sup>+</sup> cells and a strong increase in p16<sup>INK4A</sup> expression (Extended Data Fig. 8a–e). This finding suggests that IL-1α mainly plays a tumour suppressive role in *Pten*-loss-induced cellular senescence.

Different types of chemotherapy drug are known to induce senescence in tumours<sup>25</sup>. Therefore, we reasoned that chemotherapy-induced senescence could also be weakened by tumour-infiltrating Gr-1<sup>+</sup> myeloid cells. Notably, the chemokines CXCL1 and CXCL2, which act through the chemokine receptor CXCR2 to recruit Gr-1<sup>+</sup> myeloid cells<sup>8</sup>, were strongly upregulated in *Pten*<sup>PC-/-</sup> tumours (Supplementary Table 1 and Extended Data Fig. 8f). We next combined docetaxel treatment, which drives senescence in tumours<sup>22</sup>, with an antagonist of CXCR2 to block the recruitment of Gr-1<sup>+</sup> myeloid cells to *Pten*<sup>PC-/-</sup> tumours (Extended Data Fig. 8g–j). Our pre-clinical study showed that treatment with the CXCR2 antagonist synergized with docetaxel. Indeed, in mice treated with the CXCR2 antagonist and docetaxel, we observed a strong anti-tumour response and a concomitant reduction in the IL-1RA levels (Fig. 4a, b and Extended Data Fig. 9a). These changes were associated with an enhanced senescence response, reduced proliferation and the absence of apoptosis (Fig. 4c–e and Extended Data Fig. 9b–d). Notably, treating *Pten*<sup>-/-</sup> MEFs with the CXCR2 antagonist did not affect senescence (Extended Data Fig. 9e, f). Next, we assessed the levels of IL-1RA



**Figure 4 | Impaired recruitment of Gr-1<sup>+</sup> myeloid cells enhanced chemotherapy-induced senescence and chemotherapy efficacy in *Pten*<sup>PC-/-</sup> tumours: relevance for human cancer.** **a**, **b**, Tumour volume and quantification of PIN-affected glands in *Pten*<sup>PC-/-</sup> mice treated with docetaxel or a CXCR2 antagonist (CXCR2a). **c**–**e**, Ki-67 and p16<sup>INK4A</sup> immunohistochemical staining (Ki-67 and p16<sup>INK4A</sup>, brown; nuclei, blue) (c) and Ki-67 (d) and p16<sup>INK4A</sup> (e) quantification in *Pten*<sup>PC-/-</sup> mice treated with docetaxel, CXCR2a or both. Scale bar, 10 μm. **f**, CD33 and p16<sup>INK4A</sup> immunohistochemical staining in consecutive tissue microarray (n = 92) sections of human prostate cancer from two donors. CD33 and p16<sup>INK4A</sup>, brown; nuclei, blue. Main images, magnification ×400; insets, magnification ×100. **g**, p16<sup>INK4A</sup> and CD33 immunohistochemical staining in sections of PIN and prostate cancer (PCA) from the same donor. CD33 and p16<sup>INK4A</sup>, brown; nuclei, blue. Main images, magnification ×400; insets, Ki-67 positive cells in the same region at the same magnification. **a**, **b**, **d**, **e**, *n* = 5 control group; *n* = 7 treated groups; 1 tumour per mouse; 3 sections per mouse; ≥5 fields per section. Error bars, mean ± s.e.m. *P* values were derived from an unpaired, two-tailed Student's *t*-test (\**P* < 0.05; \*\**P* < 0.01; \*\*\**P* < 0.001).

in primary tumours from patients with high-risk, localized prostate cancer who received docetaxel after prostatectomy in a prospective multicentre trial<sup>26</sup>. Interestingly, patients with high levels of intratumoural IL-1RA did not respond to docetaxel and had a short disease-free survival (15 ± 10 months, mean ± s.e.m.; *P* = 0.04) compared with patients with normal IL-1RA levels (21 ± 13 months) (Extended Data Fig. 9g). Finally, we looked at the correlation between tumour-infiltrating CD33<sup>+</sup> myeloid cells and p16<sup>INK4A</sup> senescent cells in a human tissue microarray of prostate cancer (n = 92 cases) and found that the majority of tumour samples infiltrated by CD33<sup>+</sup> myeloid cells stained negative for p16<sup>INK4A</sup> (Fig. 4f and Extended Data Fig. 10a). This result was also confirmed in a panel of single human prostate cancer sections (n = 18), including areas of prostatic intraepithelial neoplasia and prostate cancer (Extended Data Fig. 10b). Moreover, we found that the majority of the prostatic intraepithelial neoplasia areas analysed had high p16<sup>INK4A</sup> staining and low CD33 and Ki-67 staining. Conversely, 90% of the prostate cancer areas stained negative for p16<sup>INK4A</sup> and positive for CD33 and Ki-67 (Fig. 4g and Extended Data Fig. 10c). These findings suggest that myeloid cells may promote tumour progression by opposing senescence in cancer in humans, in addition to mice. Bioinformatic analysis of data from the Pan-Cancer analysis project revealed that patients with low levels of *PTEN*

and high levels of *IL1RA* and *CD33* messenger RNA had a shorter survival than the other groups (Extended Data Fig. 10d).

Taken together, our data allow novel insight into the mechanisms that regulate senescence *in vivo* (Extended Data Fig. 10e). Here we provide, to our knowledge, the first evidence that an immune cell subset can antagonize senescence driven by loss of a tumour suppressor gene *in vivo*, demonstrating that senescence evasion by tumour cells can also occur in a non-cell-autonomous manner. This finding is of great relevance since senescence evasion is commonly ascribed to genetic alterations that are unlikely to occur with high frequency in senescent arrested cells or in cells committed to senescence<sup>10,27–29</sup>.

Our study supports a model whereby Gr-1<sup>+</sup> myeloid cells protect *Pten*<sup>−/−</sup> tumour cells from senescence by interfering with the balance between IL-1 $\alpha$  and IL-1RA in the tumour microenvironment (Extended Data Fig. 10f). Accordingly, treatments that block the recruitment of Gr-1<sup>+</sup> cells or decrease the levels of IL-1RA can tilt the IL-1 $\alpha$  to IL-1RA balance, reinforcing senescence in tumours and enhancing the efficacy of chemotherapy.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 14 October 2013; accepted 1 July 2014.**

**Published online 24 August; corrected online 5 November 2014 (see full-text HTML version for details).**

- Collado, M. & Serrano, M. Senescence in tumours: evidence from mice and humans. *Nature Rev. Cancer* **10**, 51–57 (2010).
- Chen, Z. *et al.* Crucial role of p53-dependent cellular senescence in suppression of *Pten*-deficient tumorigenesis. *Nature* **436**, 725–730 (2005).
- Braig, M. *et al.* Oncogene-induced senescence as an initial barrier in lymphoma development. *Nature* **436**, 660–665 (2005).
- Tasdemir, N. & Lowe, S. W. Senescent cells spread the word: non-cell autonomous propagation of cellular senescence. *EMBO J.* **32**, 1975–1976 (2013).
- Acosta, J. C. *et al.* A complex secretory program orchestrated by the inflammasome controls paracrine senescence. *Nature Cell Biol.* **15**, 978–990 (2013).
- Orjalo, A. V., Bhaumik, D., Gengler, B. K., Scott, G. K. & Campisi, J. Cell surface-bound IL-1 $\alpha$  is an upstream regulator of the senescence-associated IL-6/IL-8 cytokine network. *Proc. Natl Acad. Sci. USA* **106**, 17031–17036 (2009).
- Alimonti, A. *et al.* A novel type of cellular senescence that can be enhanced in mouse models and human tumor xenografts to suppress prostate tumorigenesis. *J. Clin. Invest.* **120**, 681–693 (2010).
- Gabrilovich, D. I., Ostrand-Rosenberg, S. & Bronte, V. Coordinated regulation of myeloid cells by tumours. *Nature Rev. Immunol.* **12**, 253–268 (2012).
- Xue, W. *et al.* Senescence and tumour clearance is triggered by p53 restoration in murine liver carcinomas. *Nature* **445**, 656–660 (2007).
- Kang, T. W. *et al.* Senescence surveillance of pre-malignant hepatocytes limits liver cancer development. *Nature* **479**, 547–551 (2011).
- Braumüller, H. *et al.* T-helper-1-cell cytokines drive cancer into senescence. *Nature* **494**, 361–365 (2013).
- Trotman, L. C. *et al.* *Pten* dose dictates cancer progression in the prostate. *PLoS Biol.* **1**, e59 (2003).
- Schaefer, B. C., Schaefer, M. L., Kappler, J. W., Marrack, P. & Kedl, R. M. Observation of antigen-dependent CD8<sup>+</sup> T-cell/dendritic cell interactions *in vivo*. *Cell. Immunol.* **214**, 110–122 (2001).
- Ruiz, E. J., Oeztuerk-Winder, F. & Ventura, J. J. A paracrine network regulates the cross-talk between human lung stem cells and the stroma. *Nature Commun.* **5**, 3175 (2014).
- Francis, K. & Palsson, B. O. Effective intercellular communication distances are determined by the relative time constants for cyto/chemokine secretion and diffusion. *Proc. Natl Acad. Sci. USA* **94**, 12258–12262 (1997).
- Ahn, G. O. & Brown, J. M. Matrix metalloproteinase-9 is required for tumor vasculogenesis but not for angiogenesis: role of bone marrow-derived myelomonocytic cells. *Cancer Cell* **13**, 193–205 (2008).
- Marigo, I. *et al.* Tumor-induced tolerance and immune suppression depend on the C/EBP $\beta$  transcription factor. *Immunity* **32**, 790–802 (2010).
- Lukacs, R. U., Goldstein, A. S., Lawson, D. A., Cheng, D. & Witte, O. N. Isolation, cultivation and characterization of adult murine prostate stem cells. *Nature Protocols* **5**, 702–713 (2010).
- Horai, R. *et al.* Production of mice deficient in genes for interleukin (IL)-1 $\alpha$ , IL-1 $\beta$ , IL-1 $\alpha/\beta$ , and IL-1 receptor antagonist shows that IL-1 $\beta$  is crucial in turpentine-induced fever development and glucocorticoid secretion. *J. Exp. Med.* **187**, 1463–1475 (1998).
- Sgroi, A. *et al.* Interleukin-1 receptor antagonist modulates the early phase of liver regeneration after partial hepatectomy in mice. *PLoS ONE* **6**, e25442 (2011).
- Tamassia, N. *et al.* Uncovering an IL-10-dependent NF- $\kappa$ B recruitment to the *IL-1ra* promoter that is impaired in STAT3 functionally defective patients. *FASEB J.* **24**, 1365–1375 (2010).
- Schwarze, S. R., Fu, V. X., Desotelle, J. A., Kenowski, M. L. & Jarrard, D. F. The identification of senescence-specific genes during the induction of senescence in prostate cancer cells. *Neoplasia* **7**, 816–823 (2005).
- Collado, M. *et al.* Tumour biology: senescence in premalignant tumours. *Nature* **436**, 642 (2005).
- Gabrilovich, D. I. & Nagaraj, S. Myeloid-derived suppressor cells as regulators of the immune system. *Nature Rev. Immunol.* **9**, 162–174 (2009).
- Ewald, J. A., Desotelle, J. A., Wilding, G. & Jarrard, D. F. Therapy-induced senescence in cancer. *J. Natl. Cancer Inst.* **102**, 1536–1546 (2010).
- Antonarakis, E. S. *et al.* An immunohistochemical signature comprising PTEN, MYC, and Ki67 predicts progression in prostate cancer patients receiving adjuvant docetaxel after prostatectomy. *Cancer* **118**, 6063–6071 (2012).
- Jacobs, J. J. *et al.* Senescence bypass screen identifies TBX2, which represses *Cdkn2a* (*p19<sup>ARF</sup>*) and is amplified in a subset of human breast cancers. *Nature Genet.* **26**, 291–299 (2000).
- Romanov, S. R. *et al.* Normal human mammary epithelial cells spontaneously escape senescence and acquire genomic changes. *Nature* **409**, 633–637 (2001).
- Berns, K. *et al.* A large-scale RNAi screen in human cells identifies new components of the p53 pathway. *Nature* **428**, 431–437 (2004).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** We thank L. Bühler for providing the *Il1ra* knockout mice, the F. Grassi laboratory and all members of the IRB animal core facility for technical assistance and the animal work, F. Stoffel for providing human samples, D. Jarrossay for helping with the flow cytometry analysis and cell sorting experiments, and all members of the Alimonti laboratory for scientific discussions. We thank C. Pissot-Soldermann, who developed NVP-BSK805. We thank T. Radimersky for providing NVP-BSK805. The human tissue microarray (as described in ref. 26) was obtained from the Department of Defense Prostate Cancer Research Program, Awards W81XWH-10-2-0056 and W81XWH-10-2-0046, Prostate Cancer Biorepository Network (PCBN). We thank G. Chiorino, I. Gregnanin, P. Ostano and L. Sacchetto for the gene expression analysis performed on myeloid and tumour cells. This work was supported by Swiss national science foundation (SNF) grant Ambizione (PZ00P3\_136612/1), the European Society for Medical Oncology (ESMO) translational research award to A.A., the Swiss Bridge Award to A.A., PEOPLE-IRG (22484), a European Research Council starting grant (ERCsg 261342), ABREOC, the Train COFUND Marie Curie to D.D.M. and Fondazione IBSA.

**Author Contributions** A.A. and A.T. originally developed the concept, further elaborated on it and designed the experiments together with D.D.M., D.D.M., A.T. and J.J.C. performed experiments and analysed the data. R.D., E.M., A.T. and D.D.M. established and carried out fluorescence microscopy. D.D.M. and T.R.J. carried out adoptive transfer experiments. M.S. and S.P. performed immunohistochemical experiments and analysis. I.G. performed experiments, R.G.-E. and C.C. carried out the bioinformatics analysis. S.D.S.-A. and M.C. provided the *K-ras*<sup>+G12V</sup> tumour samples. M.E. and Z.Z. provided tumour samples for the human prostate cancer study. A.T., D.D.M., F.G. and A.A. interpreted the data and wrote the paper. D.D.M. and A.T. contributed equally to this work.

**Author Information** The gene expression data have been deposited in Gene Expression Omnibus under accession number GSE58413. Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.A. ([andrea.alimonti@ior.ios.ch](mailto:andrea.alimonti@ior.ios.ch)).



## METHODS

**Animals.** All mice were maintained under specific-pathogen-free conditions in the animal facilities of the IRB Institute, and experiments were performed according to state guidelines and approved by the local ethics committee. Male *Pten*<sup>pc-/-</sup> and *Pten*<sup>pc+/-</sup> mice were generated and genotyped as previously described<sup>2,12</sup>. Male CByJ.B6-Tg(*UBC-GFP*)30Scha/J<sup>13</sup> transgenic mice that express GFP under the *UBC* promoter were provided by F. Grassi and were genotyped as previously described (at IRB Bellinzona)<sup>13</sup>. Male *Il1ra* knockout mice (*Il1ra*<sup>-/-</sup>) were provided by L. Bühler and were genotyped as previously described<sup>20</sup>.

For experiments involving animals, the sample size was chosen taking into consideration the means of the target values between the experimental group and the control group, the standard deviation and the statistical analysis used. For ethical reasons, the minimum number of animals necessary to achieve the scientific objectives was used. Animals were allocated randomly to each treatment group. Different treatment groups were processed identically, and animals in different treatment groups were exposed to the same environment.

**Human prostate samples.** Human tissue microarray (TMA) analyses were carried out using previously published data sets<sup>26</sup> and commercially available TMAs. Anonymized human tissue samples were obtained from the Cantonal Institute of Pathology (Locarno, Switzerland). Informed consent was obtained from all subjects. The study was approved by the Canton Ticino Ethics Committee. Three-micrometre sections were cut from formalin-fixed paraffin-embedded (FFPE) blocks and mounted on positively charged slides. Histological classification was carried out on slides stained with haematoxylin and eosin. The histological diagnosis was determined during routine pathological assessment. Slides were blindly evaluated by at least two investigators.

**Cells.** Primary MEFs were derived from littermate embryos and obtained by crossing *Pten*<sup>loxP/loxP</sup> animals as previously described<sup>27,30</sup>. Embryos were harvested at 13.5 days post coitum, and individual MEFs were produced and cultured as previously described<sup>2,12</sup>. At passage 2, the cells were harvested for western blot analysis. In the culture experiments, conditioned medium were from either *Il1ra*<sup>+/-</sup> or *Il1ra*<sup>-/-</sup> cultures. In some experiments, bone marrow precursors were collected from the femurs of *Pten*<sup>+/-</sup> mice and polarized *in vitro* in the presence of granulocyte-macrophage colony-stimulating factor (GM-CSF) and IL-6 towards the Gr-1<sup>+</sup> cell phenotype. On day 4, the conditioned medium from Gr-1<sup>+</sup> myeloid cells was collected, filtered and transferred to *Pten*<sup>-/-</sup> MEFs. Cultures were stopped 48 h later, and cells were harvested for protein extraction or stained for analysis (see also Extended Data Fig. 1m for scheme). Senescence was assessed by means of an SA- $\beta$ -gal assay (Cell Signaling Technology). Gr-1<sup>+</sup> cells were pre-treated for 24 h with NVP-BSK805 (a JAK2 inhibitor).

For all of the *in vitro* experiments, the sample size was chosen taking into consideration the means of the target values between the experimental group and the control group, the standard deviation and the statistical analysis used. Most of the *in vitro* experiments were performed in a non-blinded manner. Nevertheless, in some cases, the results were collected by an investigator other than the one who performed the experiment, to ensure blinded evaluations.

**In vitro differentiation of Gr-1<sup>+</sup> myeloid cells.** Gr-1<sup>+</sup> cells were differentiated *in vitro* as previously described<sup>17</sup>. Briefly, bone marrow precursors were flushed from the long bones of *Pten*<sup>+/-</sup> mice or *UBC-GFP* mice with RPMI 1640 medium. The cell pellet was resuspended ( $1 \times 10^6$  cells ml<sup>-1</sup>) in RPMI 1640 containing 10% heat-inactivated FBS, and the cells were cultured *in vitro* in the presence of 10 ng ml<sup>-1</sup> GM-CSF and 40 ng ml<sup>-1</sup> IL-6. On day 4, the cells were harvested and analysed by flow cytometry and qPCR (see also Extended Data Fig. 1m for scheme).

**Bone marrow transplantation.** Bone marrow was flushed from the long bones of male *Il1ra*<sup>+/-</sup> or *Il1ra*<sup>-/-</sup> mice under sterile conditions with RPMI 1640 or HBSS using a 21-gauge needle. Mononuclear cells were filtered, collected and checked for viability using trypan blue. Before transplantation, the bone marrow derived from donor mice was depleted of CD3<sup>+</sup> T cells, NK1.1<sup>+</sup> NK cells and CD19<sup>+</sup> B cells by magnetic bead separation. Recipient C57BL/6 *Pten*<sup>pc-/-</sup> mice were given 900 cGy total-body irradiation (at 7–8 weeks of age), and all mice received an eye inoculum comprising  $4.0 \times 10^6$  bone marrow cells from either *Il1ra*<sup>+/-</sup> or *Il1ra*<sup>-/-</sup> mice. Bone marrow precursors were delivered 2 h after irradiation. All mice ( $n = 4$  per group) survived (see also the scheme in Fig. 3a).

**Western blotting, immunohistochemistry and immunofluorescence.** Tissue and purified epithelial lysates were prepared with RIPA buffer (1× PBS, 1% NP-40, 0.5% sodium deoxycholate, 0.1% SDS and protease inhibitor cocktail (Roche)). The following antibodies were used for western blotting: rabbit polyclonal anti-p16<sup>INK4A</sup> (M156; Santa Cruz Biotechnology), and mouse monoclonal anti- $\beta$ -actin (AC-74; Sigma), anti-PAI1 (H-135; Santa Cruz Biotechnology), anti-IL-1RA (H-110; Santa Cruz Biotechnology) and anti-E-cadherin (36/E-cadherin; BD Biosciences). For immunohistochemistry (IHC), tissues were fixed in 10% formalin and embedded in paraffin in accordance with standard procedures. Sections were stained with anti-p53 (Accurate Chemical), anti-p16<sup>INK4A</sup> (M156; Santa Cruz Biotechnology), anti-Ki-67 (clone SP6; Lab Vision), anti-PAI1 (H-135; Santa Cruz Biotechnology) or

anti-cleaved-caspase-3 (9661, Cell Signaling Technology) antibodies. Immunofluorescence (IF) on paraffin-embedded sections was conducted with anti-vimentin (RV202; Abcam), anti-Ki-67 (clone SP6; Lab Vision) and anti-Ly-6G (Gr-1) antibodies. Confocal images were obtained with a TCS SP5 confocal microscope (Leica).

**Prostatic epithelial cell purification and cytokine array.** Eight-week-old *Pten*<sup>pc+/-</sup> and *Pten*<sup>pc-/-</sup> mice were euthanized, and whole prostates were isolated and processed to single-cell suspensions<sup>18</sup> for magnetic-activated cells sorting (MACS). Single cells were stained with FITC-anti-CD34 (stroma), FITC-anti-Ter119 (erythrocytes), FITC-anti-CD31 (endothelial cells) and FITC-anti-CD45 (leukocytes) antibodies and incubated for 20 min on ice. All antibodies (BD Biosciences) were used at 1:300; cells were then loaded onto an MS column (Miltenyi Biotec) for MACS separation, and unstained epithelial cells were collected in the negative fraction. Purified prostatic epithelial cells were processed as indicated by the manufacturer's instructions in the cytokine array kit (R&D Systems). Developed films were scanned, and the obtained images were analysed using ImageJ 1.43u; background signals were subtracted from the experimental values.

**Osmotic pump implantation.** Micro-osmotic pumps filled with PBS or recombinant IL-1 $\alpha$  ( $3 \mu\text{g kg}^{-1}$ ) were implanted in the peritoneal cavity of two groups of age-matched *Pten*<sup>pc-/-</sup> mice, to expose the prostate tissue to a continuous and controlled concentration of vehicle or protein, respectively. Briefly, mice were anaesthetized, and a midline skin incision was made in the lower abdomen. A pump was inserted into the peritoneal cavity; the muscle layer was sutured; and the skin incision was closed with wound clips or suturing.

**Autopsy and histopathology.** Animals were autopsied, and all tissues were examined regardless of their pathological status. Normal and tumour tissue samples were fixed in 10% neutral-buffered formalin (Sigma) overnight. Tissues were processed by ethanol dehydration and embedded in paraffin according to standard protocols. Sections (5  $\mu\text{m}$ ) were prepared for antibody detection and haematoxylin and eosin staining. To evaluate evidence of invasion, sections were cut at 20- $\mu\text{m}$  intervals through the tissue and stained with haematoxylin and eosin. Slides were prepared containing three to five of these sections.

**Flow cytometry analysis.** For phenotype analysis, the isolated cells were re-suspended in PBS containing 1% FCS (Sigma-Aldrich) and were pre-incubated with a purified anti-mouse CD16/CD32 antibody (eBioscience) for 30 min at room temperature. The cells were then washed and stained for 15 min at room temperature with the following anti-mouse monoclonal antibodies: CD45 eFluor 450 (clone 30-F11); Gr-1-PE (clone RB6-8C5); CD11b-APC (clone M1/70); F4/80 eFluor780 (clone BM8); NK1.1-PE (clone PK13); and CD19-FITC (clone 6D5). All of the antibodies were purchased from eBioscience. Samples were acquired on a FACSCanto II flow cytometer (BD Biosciences) after fixation with 1% formaldehyde (Sigma-Aldrich). When needed, cells were sorted from the prostate single-cell suspension using a FACSaria Cell Sorter (BD Biosciences) after staining with specific antibodies for 30 min at 4 °C in PBS containing 1% FCS. Data were analysed using FlowJo software (Tree Star).

**Treatment of mice with CXCR2 antagonist and docetaxel.** For Gr-1<sup>+</sup> myeloid cell depletion, 7-week-old *Pten*<sup>-/-</sup> mice were intraperitoneally injected with a CXCR2 antagonist (SB265610, 2 mg kg<sup>-1</sup> in sterile PBS; Tocris) once a day. For the combination treatment, docetaxel was intraperitoneally injected (10 mg kg<sup>-1</sup>) once a week for 3 weeks. Animals were killed at 10 weeks of age, and prostate tissues were harvested.

**Quantitative PCR (qPCR).** RNA isolation (QIAGEN) and TaqMan reverse transcriptase reactions (Applied Biosystems) were performed according to the manufacturer's instructions. qPCR reactions (SYBR Green system; Bio-Rad) for each sample were conducted in triplicate. The primer sequences were obtained from PrimerBank (<http://pga.mgh.harvard.edu/primerbank/index.html>). Each value was normalized to the *Gapdh* level as a reference. The primer sequences used were as follows: *Pai1* forward, 5'-TTGAATCCCATAGCTGCTT-3'; *Pai1* reverse, 5'-GACACGCCATA GGGAGAGA-3'; *p16*<sup>INK4A</sup> forward, 5'-CGCAGGTTCTTGGTCACTGT-3'; *p16*<sup>INK4A</sup> reverse, 5'-TGTTCACGAAAGCCAGCG-3'; *Il6* forward, 5'-TAGTCTTCC TACCCCAATTT-3'; *Il6* reverse, 5'-TTGGTCTTAGCACTCCTTC-3'; *Ccl2* forward, 5'-GTGGGGCGTTAACTGCAT-3'; *Ccl2* reverse, 5'-CAGGTCCCT GTCATGCTTCT-3'; *Tgfb* forward, 5'-CTCCCGTGGCTTCTAGTGC-3'; *Tgfb* reverse, 5'-GCCTTAGTTTGGACAGGATCTG-3'; *Il1ra* forward, 5'-CTGCAC TTCCACAGTCAGA-3'; *Il1ra* reverse, 5'-CTTAGCCGCTTCAGCTCTT-3'; *Il1a* forward, 5'-CGAAGACTACAGTCTGCCAT-3'; *Il1a* reverse, 5'-ATATG TGATGCCCTGGTGGT-3'; *Gapdh* forward, 5'-AGGTCGGTGTGAACGGAT TTG-3'; and *Gapdh* reverse, 5'-TGTAGACCATGTAGTTGAGGT-3'.

**Genome-wide gene expression analysis.** Total RNA was isolated from epithelial and myeloid cell populations using a miRNeasy Mini kit (QIAGEN) following the manufacturer instructions and was quantified using a NanoDrop ND-1000 Spectrophotometer (NanoDrop Technologies). RNA quality was assessed using an Agilent 2100 Bioanalyzer (Agilent Technologies). Gene expression profiling was carried out using the one-colour labelling method. Labelling, hybridization, washing and slide scanning were performed following the manufacturer's protocols. Briefly, equal



amounts of total RNA (100 ng) were amplified, labelled with Cy3 and purified with spin columns. Labelled specimens (600 ng) were hybridized to SurePrint G3 Mouse GE 8×60K Gene Expression Microarrays (Agilent). After 17 h, slides were washed and scanned (G2505C scanner, Agilent Technologies).

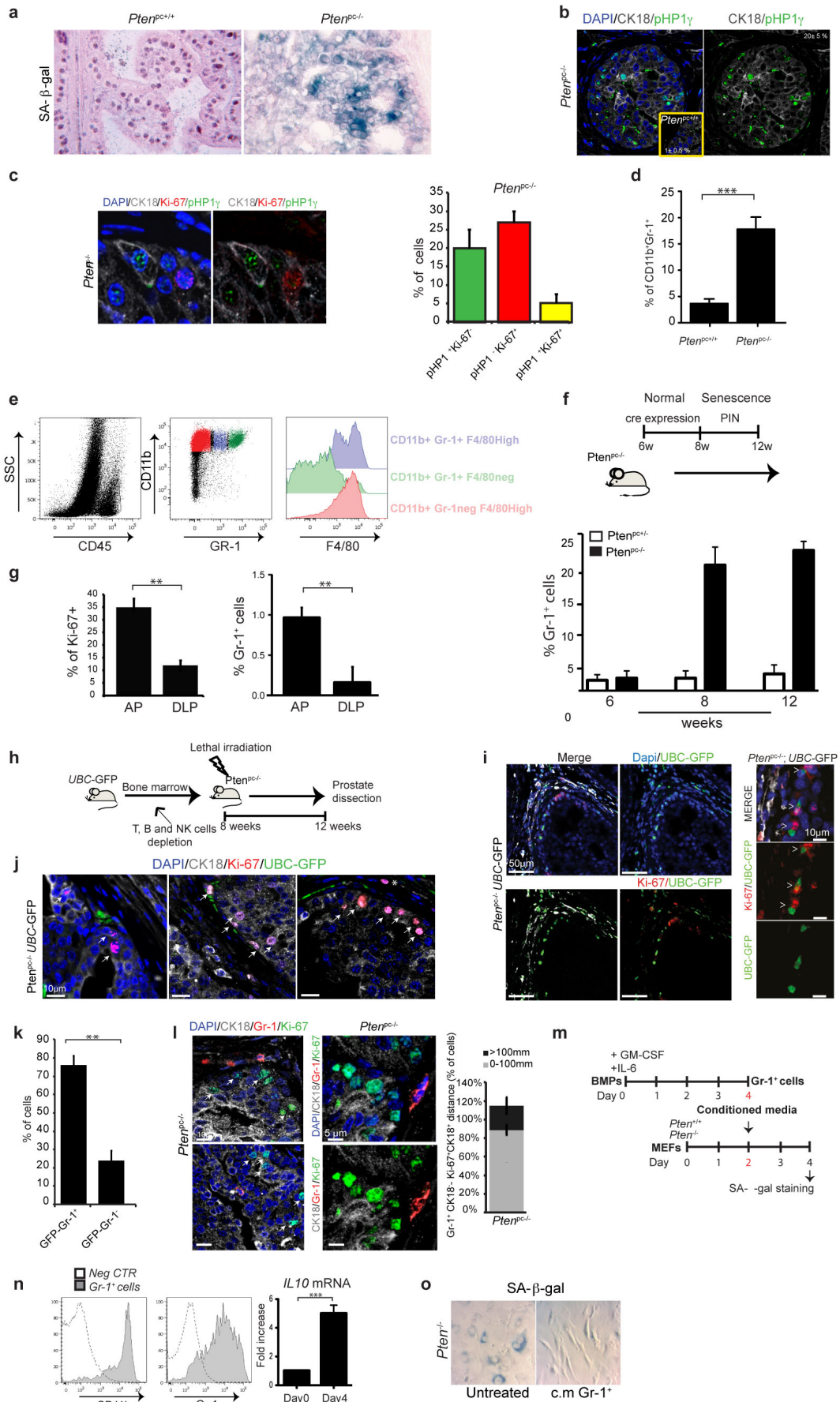
**Gene expression data analysis.** Images were analysed using Feature Extraction software v10.7 (Agilent). Raw data elaboration was carried out with Bioconductor (<http://www.bioconductor.org>), using the R (v3.0.2) statistical environment. Background correction was performed with the normexp method with an offset of 50, and quantile was used for between-array normalization. The LIMMA (Linear Models for Microarray Analysis) package was then used to identify differentially expressed genes, using the empirical Bayes method to compute a moderated *t*-statistic. Gene set enrichment analysis (GSEA; v2.07) was performed to examine the association between predefined gene sets and gene expression profiles of selected samples.

**Survival curves.** Differential survival between patient subgroups was plotted and calculated using Kaplan–Meier curves. Patients were stratified based on *IL1RA* and *CD33* score values. Briefly, scores were rank ordered and divided into seven percentiles

(from the lowest to the highest values). Such stratification showed significant differences in overall survival within The Cancer Genome Atlas Pan-Cancer analysis project (log-rank test). The Pan-Cancer data set matrix and clinical information<sup>31</sup> were downloaded from the UCSC Cancer Genomics Browser (<https://genome-cancer.ucsc.edu/>).

**Statistical analysis.** Statistical analysis of the data was performed using a two-tailed, unpaired Student's *t*-test. Values are expressed as mean ± s.e.m. (\**P* < 0.05; \*\**P* < 0.01; \*\*\**P* < 0.001). Significant differences in survival curves were calculated using the log-rank test. Correlation analysis in TMA staining evaluation was conducted with Fisher's exact test, using the estimated percentage of positively stained cells as determined by a pathologist (M.S.).

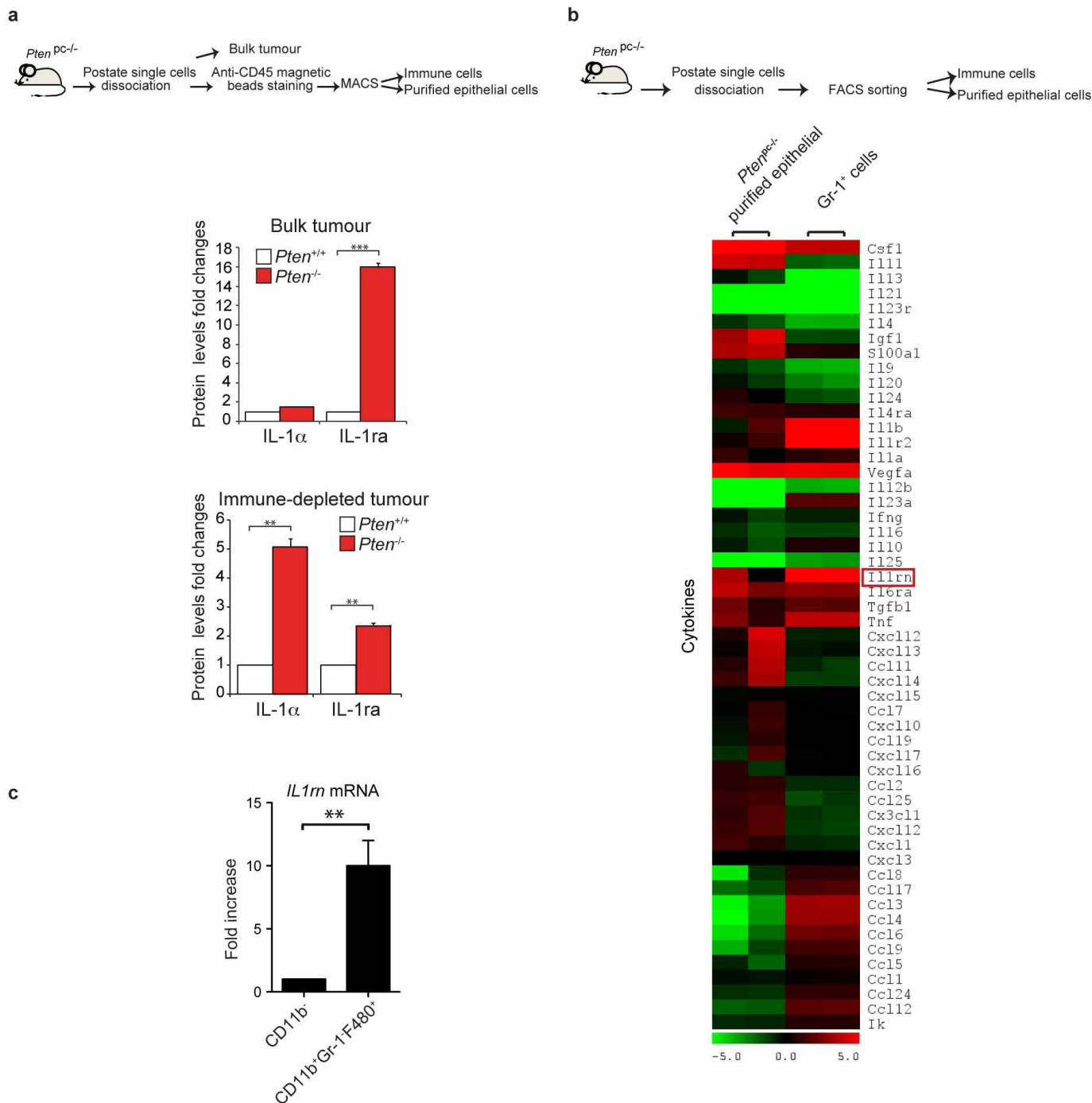
30. Alimonti, A. *et al.* Subtle variations in Pten dose determine cancer susceptibility. *Nature Genet.* **42**, 454–458 (2010).
31. The Cancer Genome Atlas Research Network. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genet.* **45**, 1113–1120 (2013).



**Extended Data Figure 1 | Gr-1<sup>+</sup> myeloid cells infiltrate senescent tumours in *Pten*<sup>pc-/-</sup> mice.** **a**, Representative SA- $\beta$ -gal staining of prostate sections from 8-week-old *Pten*<sup>pc+/+</sup> and *Pten*<sup>pc-/-</sup> mice. Original magnification,  $\times 400$ . **b**, Representative confocal immunofluorescence (IF) images showing staining of the epithelial marker cytokeratin 18 (CK18) (grey) and the senescence marker pHP1 $\gamma$  (green) in prostate tumours from *Pten*<sup>pc-/-</sup> mice. Cells were counterstained with the nuclear marker DAPI (blue). A *Pten*<sup>pc+/+</sup> prostate negative for pHP1 $\gamma$  staining is also shown (inset). **c**, Representative confocal IF image showing proliferating epithelial cells (CK18, grey; Ki-67, red) and senescent epithelial cells (cytokeratin 18, grey; pHP1 $\gamma$ , green) in *Pten*<sup>pc-/-</sup> prostate lesions. Cells were counterstained with the nuclear marker DAPI (blue). The histogram shows the quantification of CK18<sup>+</sup>pHP1 $\gamma$ <sup>+</sup>Ki-67<sup>-</sup>, CK18<sup>+</sup>pHP1 $\gamma$ <sup>-</sup>Ki-67<sup>+</sup> and CK18<sup>+</sup>pHP1 $\gamma$ <sup>+</sup>Ki-67<sup>+</sup> cells ( $n = 3$ ; 1 tumour per mouse; 10 fields acquired; 412 cells counted). **d**, Quantification of CD11b<sup>+</sup>Gr-1<sup>+</sup> immune cells in 8-week-old *Pten*<sup>pc+/+</sup> and *Pten*<sup>pc-/-</sup> mice ( $n = 6$ ). **e**, Flow cytometry analysis showing the heterogeneity of the tumour-infiltrating CD45<sup>+</sup>CD11b<sup>+</sup>Gr-1<sup>+</sup> immune cells in *Pten*<sup>pc-/-</sup> prostates. **f**, In *Pten*<sup>pc-/-</sup> tumours, the senescence response starts at 8 weeks of age (top). A time course experiment is shown, indicating the recruitment of Gr-1<sup>+</sup> myeloid cells in *Pten*<sup>pc+/+</sup> and *Pten*<sup>pc-/-</sup> mice at the onset of tumorigenesis (bottom) ( $n = 3$  per group; 1 tumour per mouse). **g**, Correlation between Ki-67 staining and percentage of Gr-1<sup>+</sup> myeloid cells in the anterior (AP) and dorsolateral lobes (DLP) of *Pten*<sup>pc-/-</sup> tumours ( $n = 3$ ; 1 tumour per mouse). **h**, Experimental scheme. *Pten*<sup>pc-/-</sup> mice were lethally irradiated and then transferred with bone marrow from *UBC-GFP* mice that had been depleted of T-, B- and natural killer (NK) cells. Prostate tissues were collected

4 weeks after transfer. **i**, Representative confocal IF images showing the localization of myeloid cells (green) infiltrating the anterior prostate gland of *Pten*<sup>pc-/-</sup>*UBC-GFP* mice. Proliferating cells (Ki-67, red) and stroma (vimentin, grey) are also shown. Cells were counterstained with DAPI (blue). **j**, Representative confocal IF image showing the localization of tumour-infiltrating *UBC-GFP* cells and proliferating epithelial cells (CK18, grey; Ki-67, red) in prostate lesions from *Pten*<sup>pc-/-</sup>*UBC-GFP* mice. Cells were counterstained with the nuclear marker DAPI (blue). Arrows indicate CK18<sup>+</sup>Ki-67<sup>+</sup> cells, which were considered for the analysis, while \* indicates CK18<sup>+</sup>Ki-67<sup>+</sup> cells, which were excluded from the analysis. **k**, Quantification of *UBC-GFP*<sup>+</sup>Gr-1<sup>+</sup> cells ( $n = 4$ ; 1 tumour per mouse; 5 fields acquired; 300 cells counted). **l**, Representative confocal IF image showing the localization of tumour-infiltrating myeloid cells (Gr-1, red) and proliferating epithelial cells (CK18, grey; Ki-67, green) in prostate lesions from non-irradiated *Pten*<sup>pc-/-</sup> mice. Cells were counterstained with the nuclear marker DAPI (blue). The histogram shows the quantification of the distance between tumour-infiltrating Gr-1<sup>+</sup>CK18<sup>-</sup> myeloid cells and CK18<sup>+</sup>Ki-67<sup>+</sup> proliferating epithelial cells ( $n = 3$ ; 10 fields acquired; 334 measurements). The arrows indicate CK18<sup>+</sup>Ki-67<sup>+</sup> cells, which were considered for the analysis. **m**, Experimental set-up. **n**, Flow cytometry and qRT-PCR analysis of Gr-1<sup>+</sup> myeloid cells differentiated *in vitro* in the presence of granulocyte-macrophage colony-stimulating factor (GM-CSF) and IL-6, showing upregulation of Gr-1 and *Il10* mRNA. **o**, SA- $\beta$ -gal staining of *Pten*<sup>-/-</sup> MEFs. **c, d, f, g, k, l, n**, Error bars, mean  $\pm$  s.e.m. *P* values were derived from an unpaired, two-tailed Student's *t*-test (\*\**P* < 0.01; \*\*\**P* < 0.001).

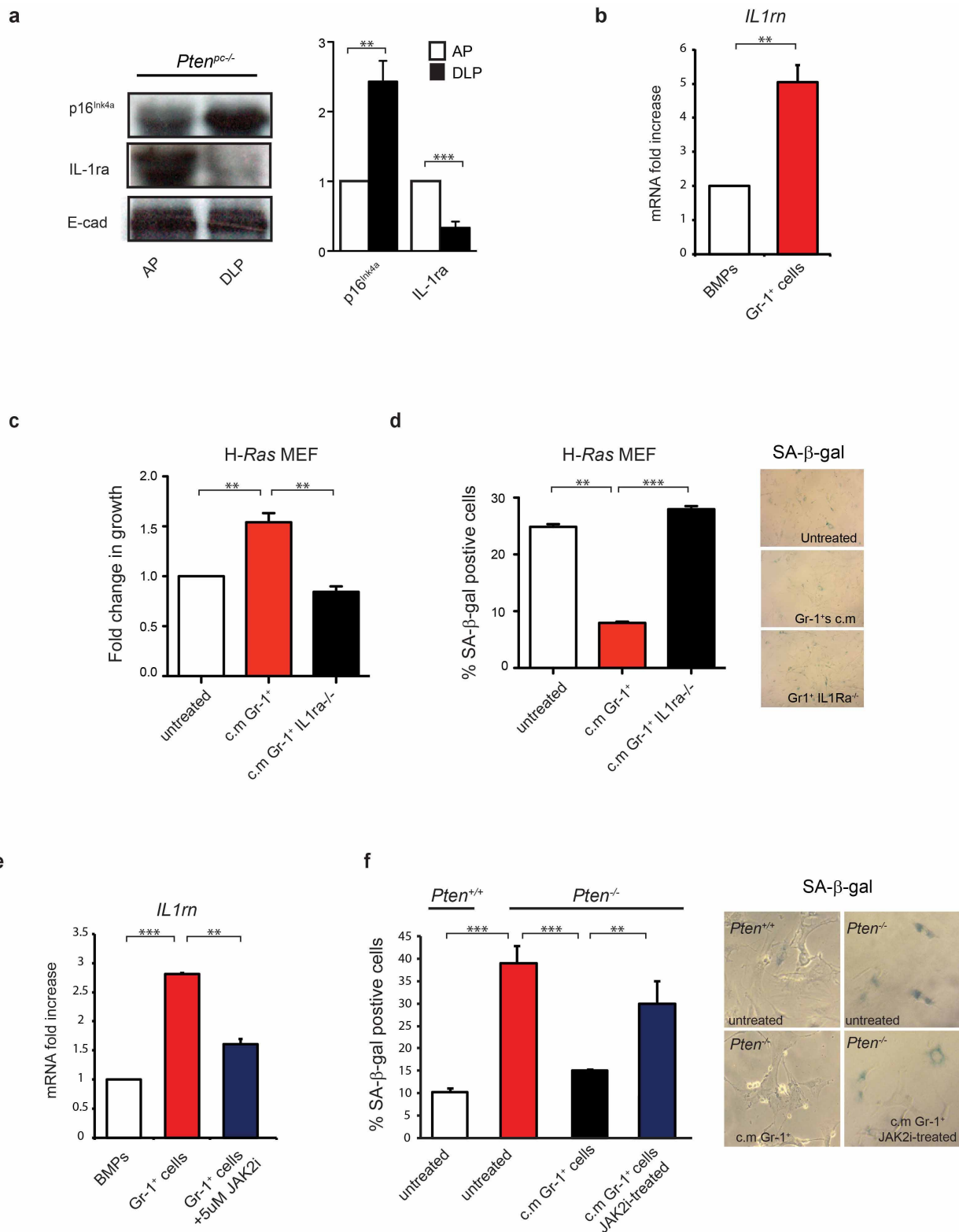




# Extended Data Figure 2 | Gene expression analysis of factors expressed by Gr-1<sup>+</sup> myeloid cells and epithelial cells sorted from *Pten*<sup>PC-/-</sup> tumours.

**a**, Experimental set-up (top). Protein levels of IL-1RA in *Pten*<sup>PC-/-</sup> bulk prostate tumours and *Pten*<sup>PC-/-</sup> immunodepleted prostatic epithelial cells (bottom) ( $n = 3$  per group; 1 tumour per mouse). **b**, Experimental set-up. Gene expression analysis of epithelial cells and Gr-1<sup>+</sup> myeloid cells purified from *Pten*<sup>PC-/-</sup> prostate tumours. Briefly, prostates were isolated from 8-week-old *Pten*<sup>PC-/-</sup> mice and processed to a single-cell suspension. CD45<sup>+</sup> epithelial cells and CD45<sup>+</sup>CD11b<sup>+</sup>Gr-1<sup>+</sup> myeloid cells were further sorted using a

FACSaria Cell Sorter. Total RNA was isolated from the epithelial and myeloid cell populations, and gene expression profiling was carried out using the one-colour labelling method, performing two replicates for each condition. A heatmap displaying the mRNA expression of 53 secreted factors is shown ( $n = 2$  per group). **c**, qRT-PCR analysis of CD11b<sup>+</sup>Gr-1<sup>+</sup>F4/80<sup>+</sup> sorted from *Pten*<sup>PC-/-</sup> prostate tumours, showing *Il1ra* expression ( $n = 3$ ). **a**, **c**, Error bars, mean  $\pm$  s.e.m.  $P$  values were derived from an unpaired, two-tailed Student's  $t$ -test (\*\* $P < 0.01$ ; \*\*\* $P < 0.001$ ).

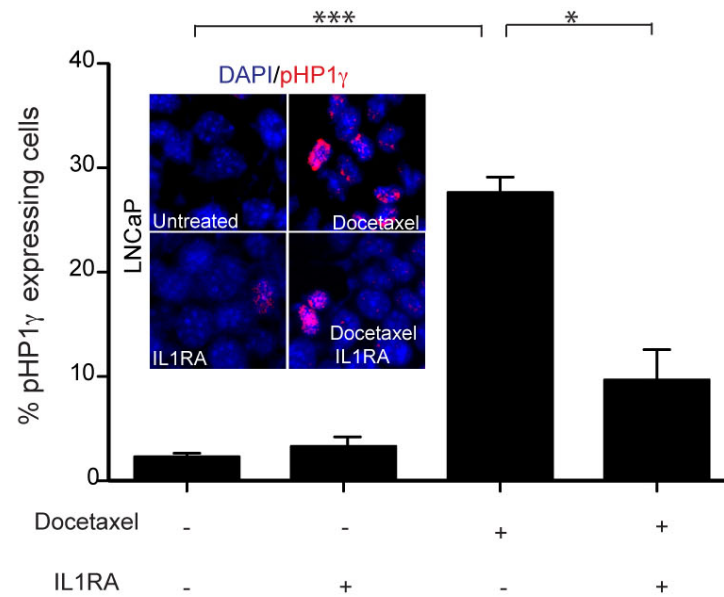


**Extended Data Figure 3 | Gr-1<sup>+</sup> myeloid cells oppose senescence in both *Pten*-loss-induced cellular senescence and oncogene-induced senescence.**

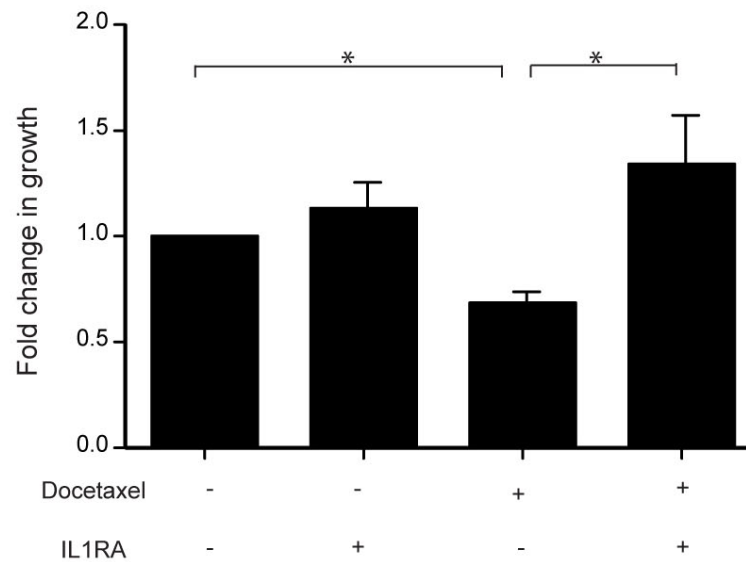
**a**, Western blot analysis showing the inverse correlation between IL-1RA and p16<sup>INK4A</sup> protein levels in the anterior (AP) and dorsolateral lobes (DLP) of *Pten*<sup>pc-/-</sup> prostate tumours. Levels are normalized to E-cadherin expression. **b**, *Il1ra* mRNA expression of bone marrow precursors (BMPs) and Gr-1<sup>+</sup> myeloid cells sorted from *Pten*<sup>pc-/-</sup> prostate tumours (*n* = 3). **c**, Cell growth of H-ras MEFs cultured in the presence of conditioned medium from Gr-1<sup>+</sup> myeloid cells (*n* = 3). **d**, Quantification (left) and representative images (right) of SA-β-gal<sup>+</sup> H-ras MEFs cultured in the presence of conditioned medium

from Gr-1<sup>+</sup> myeloid cells (*n* = 3). **e**, *Il1ra* mRNA expression of Gr-1<sup>+</sup> myeloid cells differentiated for 4 days with IL-6 and GM-CSF, in the absence or presence of the JAK2 inhibitor NVP-BSK805, compared with bone marrow precursors (BMPs) (*n* = 3). **f**, Quantification (left) and representative images (right) of SA-β-gal<sup>+</sup> *Pten*<sup>-/-</sup> MEFs cultured in the presence of conditioned medium from Gr-1<sup>+</sup> myeloid cells that had been pre-treated with the JAK2 inhibitor NVP-BSK805 (*n* = 3). **a–f**, Error bars, mean ± s.e.m. *P* values were derived from an unpaired, two-tailed Student's *t*-test (\*\**P* < 0.01; \*\*\**P* < 0.001).

**a**

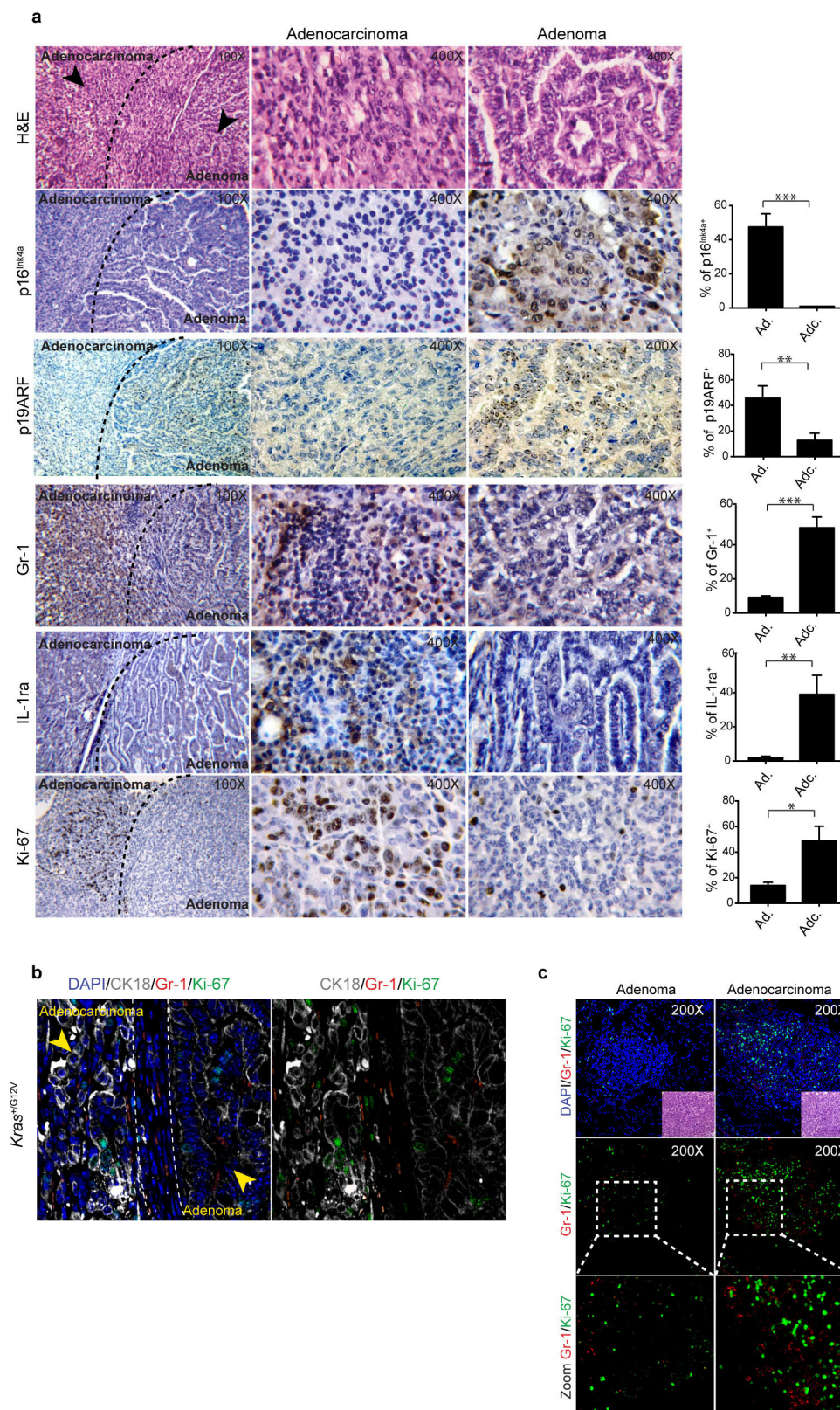


**b**



**Extended Data Figure 4 | IL-1RA opposes docetaxel-induced senescence in LNCaP cancer cells *in vitro*.** **a**, Histogram showing the quantification of pHP1 $\gamma$ <sup>+</sup> cells. Briefly, LNCaP prostate cancer cells were cultured in the absence or presence of docetaxel, with or without human recombinant IL-1RA. After 5 days, cells were collected and stained for immunofluorescence analysis. Representative confocal IF staining showing senescent pHP1 $\gamma$ <sup>+</sup> (red) LNCaP

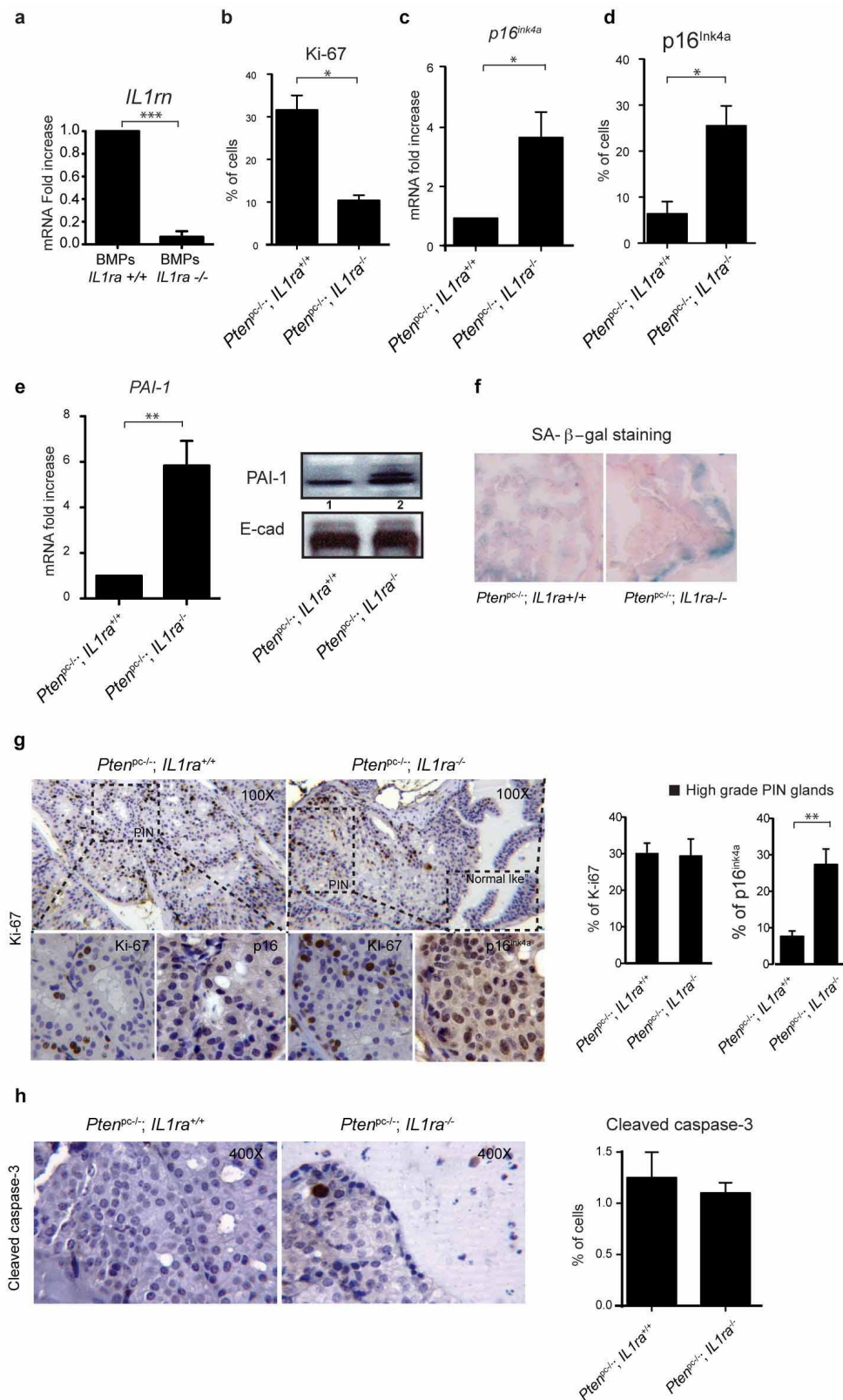
cancer cells (inset). Cells were counterstained with the nuclear marker DAPI (blue). **b**, Cell growth of LNCaP cells cultured in the absence or presence of docetaxel, with or without human recombinant IL-1RA ( $n = 3$ ). **a**, **b**, Error bars, mean  $\pm$  s.e.m.  $P$  values were derived from an unpaired, two-tailed Student's  $t$ -test (\* $P < 0.01$ ; \*\*\* $P < 0.001$ ).



**Extended Data Figure 5 | Gr-1<sup>+</sup> myeloid cells infiltrate adenocarcinoma areas in lungs from *K-ras*<sup>G12V</sup> mice.** **a**, Haematoxylin and eosin (H&E), p16<sup>INK4A</sup>, p19<sup>ARF</sup>, Gr-1, IL-1RA and Ki-67 immunohistochemical staining in lungs from *K-ras*<sup>G12V</sup> mice. Original magnification,  $\times 400$ . Staining of both adenocarcinoma and adenoma areas is shown (left). Histograms showing quantification of cells positive for p16<sup>INK4A</sup>, p19<sup>ARF</sup>, Gr-1, IL-1RA and Ki-67 (right) ( $n = 7$ ; 3 sections per mouse;  $\geq 5$  fields per section analysed). **b**, **c**, Representative confocal IF images showing staining of Ki-67 (green)

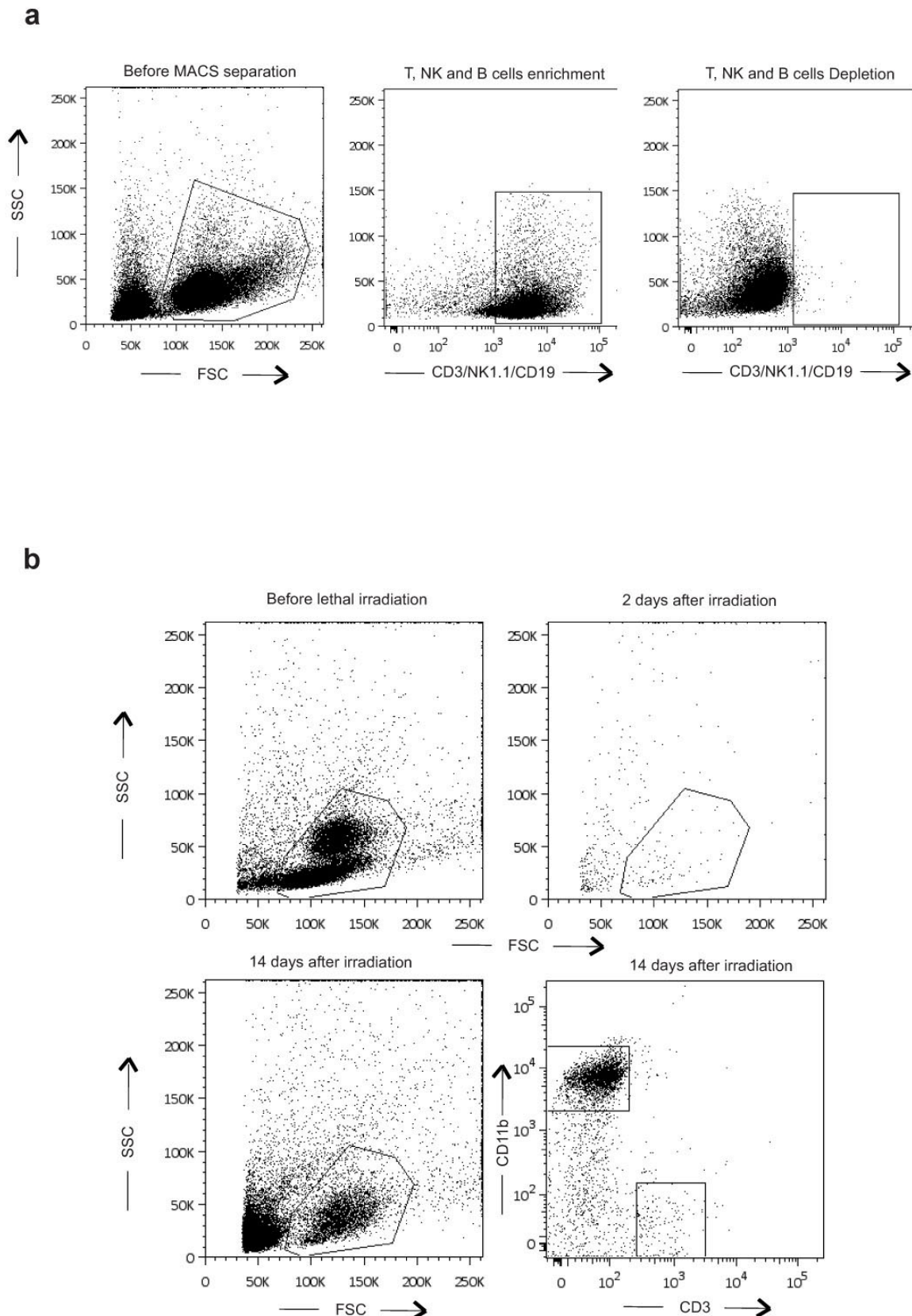
and the myeloid marker Gr-1 (red) with **(b)** or without **(c)** the epithelial marker CK18 (grey), in adenocarcinoma and adenoma areas of lungs from *K-ras*<sup>G12V</sup> mice. Cells were counterstained with the nuclear marker DAPI (blue). Panel **b** magnification  $\times 400$ . Panel **c** magnification  $\times 200$ . Top panel insets in **c** show H&E staining (magnification  $\times 100$ ) of the same areas stained for IF; bottom panels, magnification  $\times 400$ . **a**, Error bars, mean  $\pm$  s.e.m.  $P$  values were derived from an unpaired, two-tailed Student's  $t$ -test (\* $P < 0.05$ ; \*\* $P < 0.01$ ; \*\*\* $P < 0.001$ ).





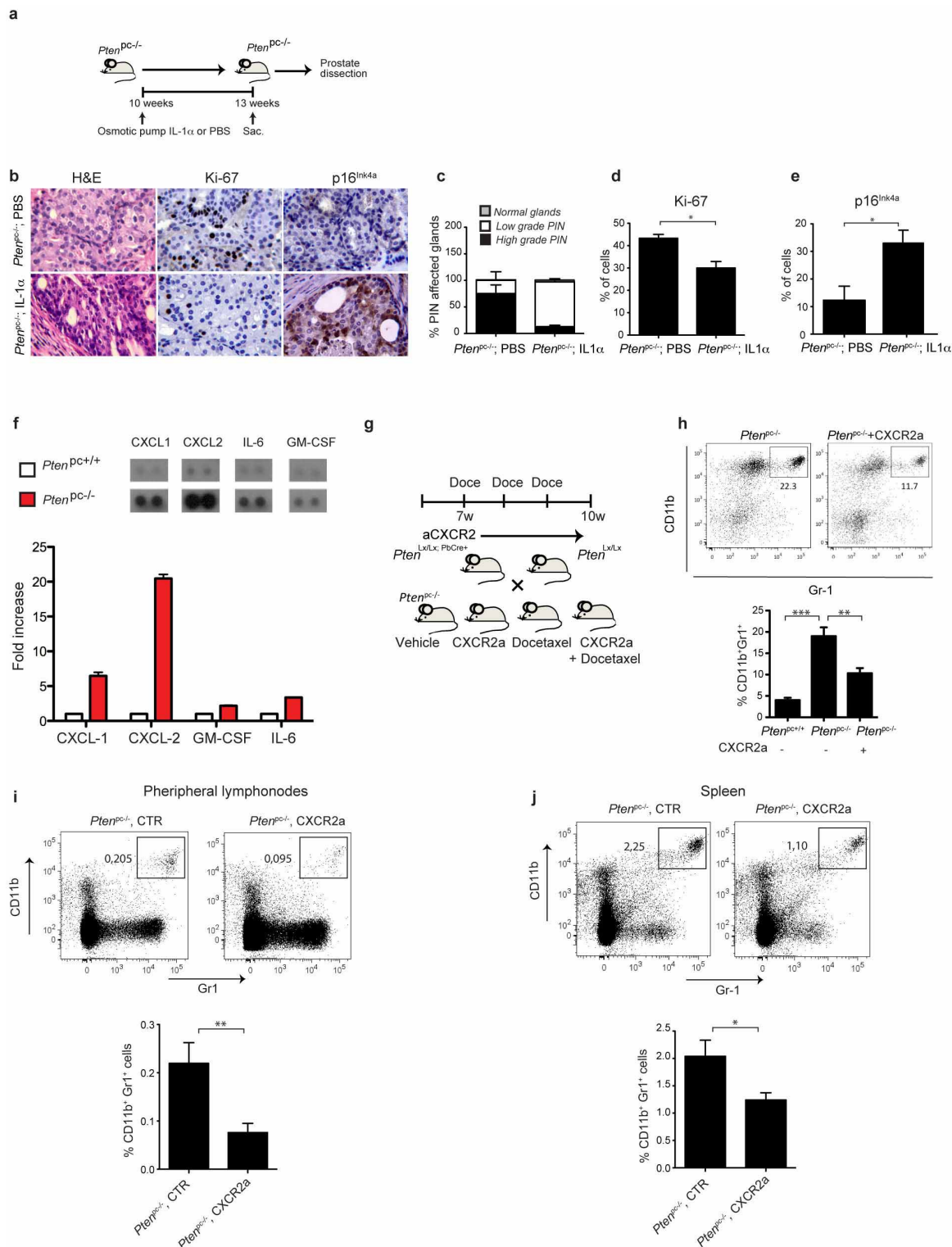
**Extended Data Figure 6 | Senescence and apoptotic markers in *Pten*<sup>pc+/+</sup> *Il1ra*<sup>+/+</sup> and *Pten*<sup>pc-/-</sup> *Il1ra*<sup>-/-</sup> mice.** **a**, mRNA levels of *Il1ra* in BMPs from the indicated genotypes ( $n = 3$ ). **b**, Quantification of Ki-67 staining. **c**, **d**, *p16*<sup>INK4A</sup> mRNA and protein levels. **e**, *PAI1* mRNA and protein levels. **f**, SA-β-gal staining in prostate tissues from *Pten*<sup>pc+/+</sup> *Il1ra*<sup>+/+</sup> and *Pten*<sup>pc-/-</sup> *Il1ra*<sup>-/-</sup> mice ( $n = 4$  mice per group; 1 tumour per mouse; 3 sections per tumour;  $\geq 5$  fields per section). **g**, Ki-67 and *p16*<sup>INK4A</sup> immunohistochemical staining of stage-matched prostate tumours from

*Pten*<sup>pc+/+</sup> *Il1ra*<sup>+/+</sup> and *Pten*<sup>pc-/-</sup> *Il1ra*<sup>-/-</sup> mice. The histograms show quantification of Ki-67 and *p16*<sup>INK4A</sup> positivity ( $n = 4$  mice per group; 1 tumour per mouse; 3 sections per tumour; 3 fields per section). **h**, Immunohistochemistry for cleaved caspase-3 in prostate tissues from *Pten*<sup>pc+/+</sup> *Il1ra*<sup>+/+</sup> and *Pten*<sup>pc-/-</sup> *Il1ra*<sup>-/-</sup> mice. **a–e**, **g**, **h**, Error bars, mean  $\pm$  s.e.m.  $P$  values were derived from an unpaired, two-tailed Student's  $t$ -test (\* $P < 0.05$ ; \*\* $P < 0.01$ ; \*\*\* $P < 0.001$ ).



**Extended Data Figure 7 | Efficiency of magnetic-activated cell sorting (MACS) purification and bone marrow transplantation.** **a**, Representative flow cytometry plots showing whole bone marrow cells, phycoerythrin (PE)-positive cells isolated after magnetic separation and bone marrow cells depleted of T, NK and B cells before adoptive transfer to irradiated *Pten<sup>pc-/-</sup>* mice (gating on total cells). Briefly, cells were flushed from the long bones of donor

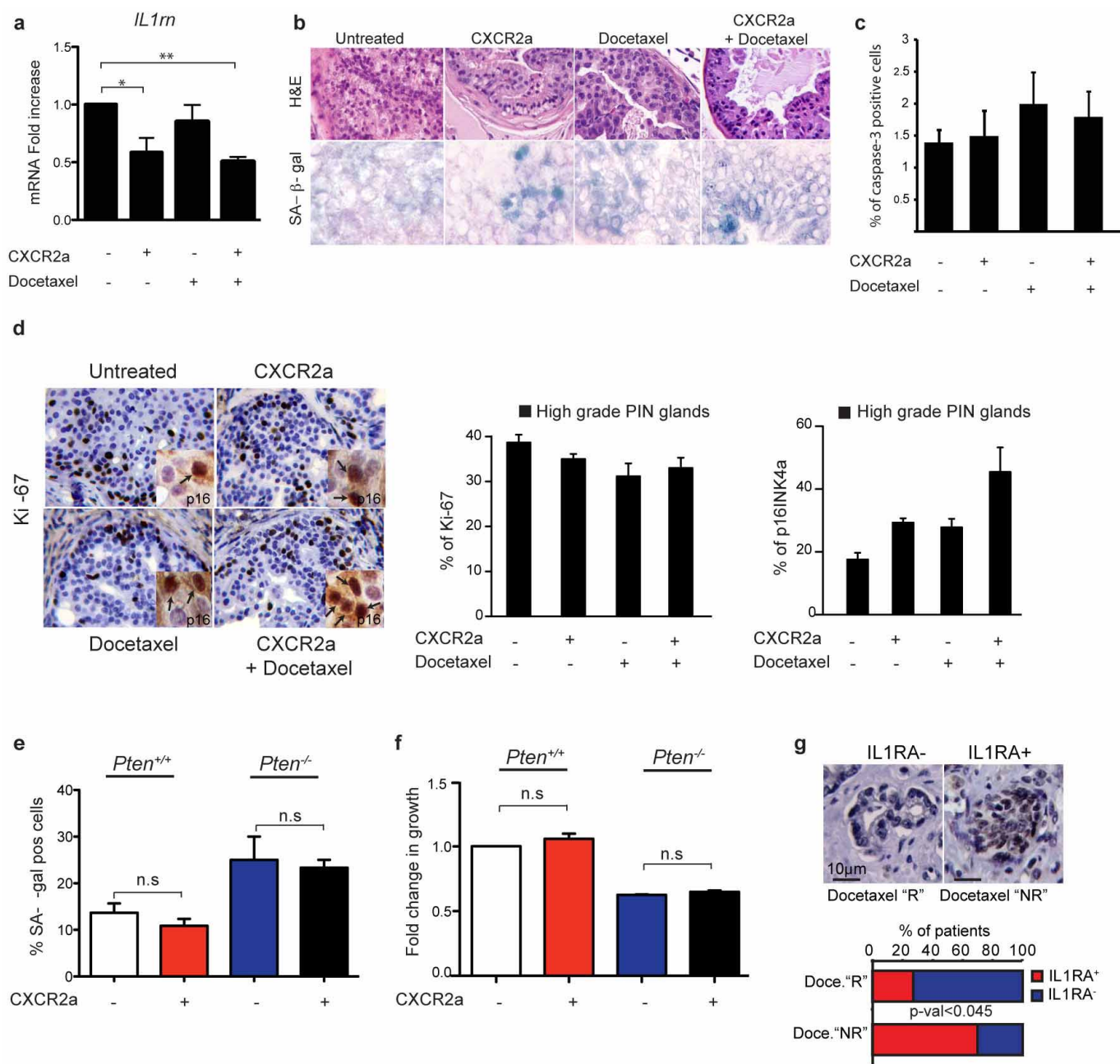
mice and stained with the following anti-mouse antibodies: anti-CD3-PE, anti-NK1.1-PE and anti-CD19-PE. The cells were then washed and stained with anti-PE magnetic beads and collected for magnetic separation. **b**, Representative plots obtained from flow cytometry analysis of splenocytes isolated from *Pten<sup>pc-/-</sup>* mice before and after lethal irradiation (top). Immune reconstitution 14 days after bone marrow transplantation (bottom).



**Extended Data Figure 8 | Effect of IL-1 $\alpha$  and CXCR2a on *Pten*<sup>pc-/-</sup> tumours.** **a**, Experimental scheme. Briefly, osmotic pumps were implanted in the peritoneal cavity of six *Pten*<sup>pc-/-</sup> mice, to expose the prostate tissue to a continuous and controlled concentration of either IL-1 $\alpha$  or PBS. **b**, Immunohistochemical staining (H&E, Ki-67 and p16<sup>INK4A</sup>) of prostate sections from *Pten*<sup>pc-/-</sup> mice treated with IL-1 $\alpha$  or PBS. **c–e**, Histograms showing quantification of glands affected by prostatic intraepithelial neoplasia (PIN) (**c**), Ki-67 positivity (**d**) and p16<sup>INK4A</sup> positivity (**e**) ( $n = 3$  per group; 1 tumour per mouse; 3 sections per mouse; 3 fields per sections were analysed). **f**, Protein profile of immunodepleted epithelial cells showing the high levels of cytokines that recruit (CXCL1 and CXCL2) and activate (GM-CSF and IL-6) Gr-1<sup>+</sup> myeloid cells in *Pten*<sup>pc-/-</sup> prostate tumours. **g**, Experimental set-up.

Doce, docetaxel. **h**, Flow cytometry plots showing the reduced recruitment of Gr-1<sup>+</sup> myeloid cells in *Pten*<sup>pc-/-</sup> mice after treatment with a CXCR2 antagonist (CXCR2a), with gating on live CD45<sup>+</sup> cells. The histogram shows the frequency of Gr-1<sup>+</sup> myeloid cells ( $n = 5$  control group;  $n = 7$  treated groups). **i**, **j**, Flow cytometry plots showing the recruitment of Gr-1<sup>+</sup> myeloid cells to the peripheral lymph nodes (upper panels) and spleen (lower panel) isolated from *Pten*<sup>pc-/-</sup> mice, after treatment with CXCR2a, with gating on live cells. The histograms (right) show the frequency of Gr-1<sup>+</sup> myeloid cells in the lymph nodes and spleen from *Pten*<sup>pc-/-</sup> mice, after treatment with CXCR2a ( $n = 5$  mice per group). **c–f**, **h–j**, Error bars, mean  $\pm$  s.e.m.  $P$  values were derived from an unpaired, two-tailed Student's  $t$ -test (\* $P < 0.05$ ; \*\* $P < 0.01$ ; \*\*\* $P < 0.001$ ).



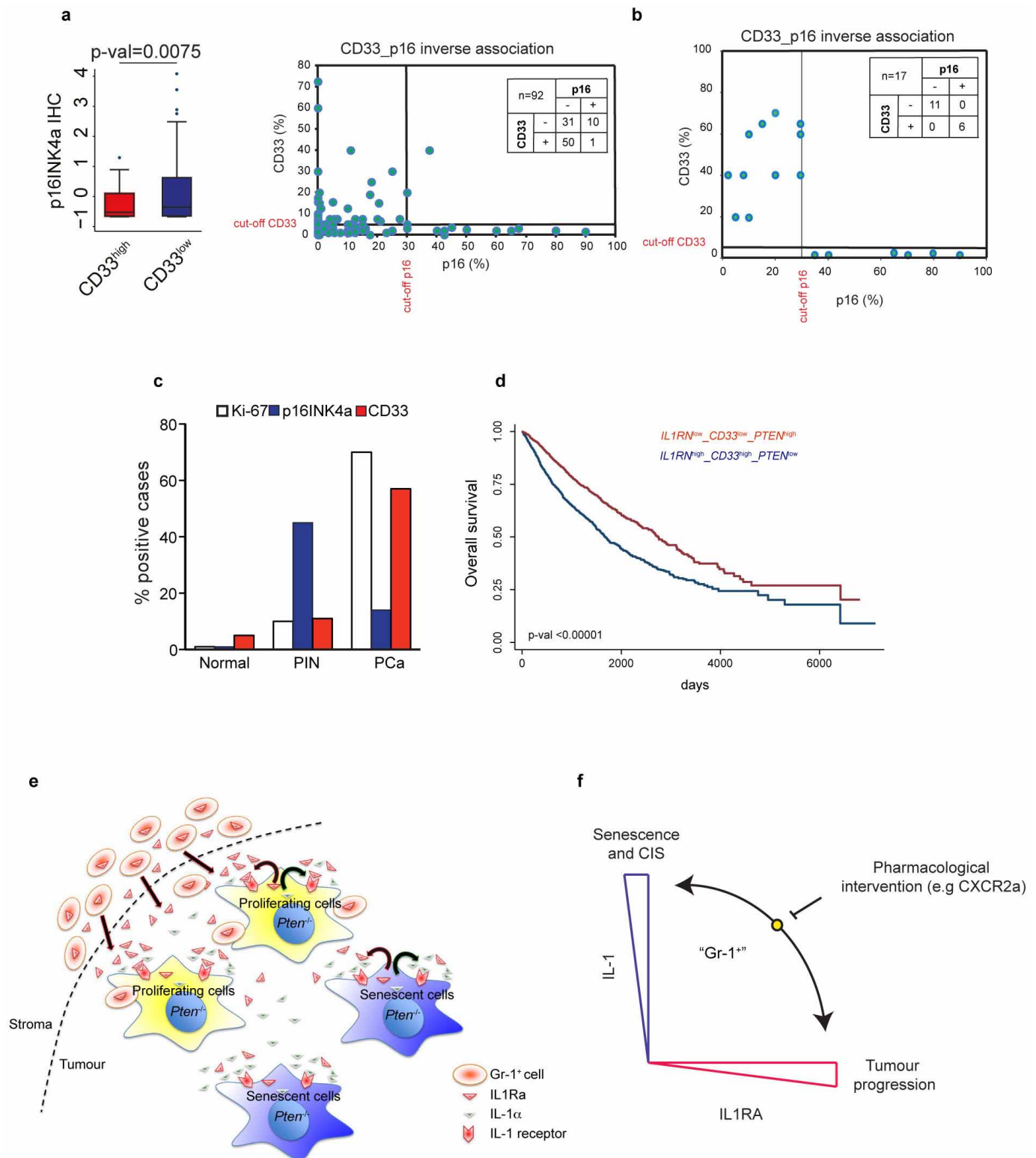


# Extended Data Figure 9 | Treatment with a CXCR2 antagonist *in vivo*.

**a**, *Il1ra* mRNA levels in *Pten*<sup>pc-/-</sup> tumours after treatment with CXCR2a alone or in combination with docetaxel. **b**, Immunohistochemical staining (H&E and SA-β-gal) in mice treated with CXCR2a and docetaxel. **c**, Quantification of cleaved caspase-3 in *Pten*<sup>pc-/-</sup> tumours after the indicated treatments. **d**, Immunohistochemical staining for Ki-67 in stage-matched prostate tumours from *Pten*<sup>pc-/-</sup> mice after treatment. The histograms show quantification

of Ki-67 and p16<sup>INK4A</sup> positivity. **e**, **f**, Treatment of *Pten*<sup>-/-</sup> MEFs with CXCR2a (*n* = 3). NS, not significant. **g**, Staining and quantification of IL-1RA in primary tumours from patients. Responder patients ("R") and non-responder patients ("NR"), based on disease-free survival. **a**, **c**–**e**, Error bars, mean ± s.e.m. *P* values were derived from an unpaired, two-tailed Student's *t*-test (\**P* < 0.05; \*\**P* < 0.01). **b**–**d**, Control *n* = 5; treated *n* = 7; 3 sections per mouse; 5 fields per section.





**Extended Data Figure 10 | Evidence in human samples and proposed model.** **a, b,** Graphs showing the inverse association between p16<sup>INK4A</sup> and CD33 in the tissue microarrays and single prostate sections from human prostate cancer. Box plots in **a** show the interquartile range, whiskers show the full range. **c,** Histogram showing the percentage of cases positive for Ki-67, p16<sup>INK4A</sup> and CD33 in sections. Normal-like prostate areas were compared with PIN and prostate cancer (PCA) areas in the same section. **d,** Kaplan–Meier analysis (see the Survival curves subsection in Methods).

**e,** Gr-1<sup>+</sup> myeloid cells recruited to the tumour site oppose *Pten*-loss-induced cellular senescence by secreting IL-1RA in the tumour microenvironment. **f,** Gr-1<sup>+</sup> myeloid cells can protect tumour cells from senescence by tilting the balance between IL-1α and IL-1RA in the tumour microenvironment. Pharmacological interventions aimed at impairing Gr-1<sup>+</sup> myeloid cell recruitment (for example, CXCR2a) can enhance senescence, thus improving chemotherapy efficacy. CIS, chemotherapy induced senescence. **a,** Correlation assessed with Fisher's exact test.

# Broad and potent HIV-1 neutralization by a human antibody that binds the gp41–gp120 interface

Jinghe Huang<sup>1</sup>, Byong H. Kang<sup>1</sup>, Marie Pancera<sup>2</sup>, Jeong Hyun Lee<sup>3,4</sup>, Tommy Tong<sup>5</sup>, Yu Feng<sup>4</sup>, Hiromi Imamichi<sup>1</sup>, Ivelin S. Georgiev<sup>2</sup>, Gwo-Yu Chuang<sup>2</sup>, Aliaksandr Druz<sup>2</sup>, Nicole A. Doria-Rose<sup>2</sup>, Leo Laub<sup>1</sup>, Kwinten Sliepen<sup>6</sup>, Marit J. van Gils<sup>6</sup>, Alba Torrents de la Peña<sup>6</sup>, Ronald Derking<sup>6</sup>, Per-Johan Klasse<sup>7</sup>, Stephen A. Migueles<sup>1</sup>, Robert T. Bailer<sup>2</sup>, Munir Alam<sup>8</sup>, Pavel Pugach<sup>7</sup>, Barton F. Haynes<sup>8</sup>, Richard T. Wyatt<sup>3,4</sup>, Rogier W. Sanders<sup>6,7</sup>, James M. Binley<sup>5</sup>, Andrew B. Ward<sup>3,4</sup>, John R. Mascola<sup>2</sup>, Peter D. Kwong<sup>2</sup> & Mark Connors<sup>1</sup>

**The isolation of human monoclonal antibodies is providing important insights into the specificities that underlie broad neutralization of HIV-1 (reviewed in ref. 1). Here we report a broad and extremely potent HIV-specific monoclonal antibody, termed 35O22, which binds a novel HIV-1 envelope glycoprotein (Env) epitope. 35O22 neutralized 62% of 181 pseudoviruses with a half-maximum inhibitory concentration (IC<sub>50</sub>) < 50 µg ml<sup>-1</sup>. The median IC<sub>50</sub> of neutralized viruses was 0.033 µg ml<sup>-1</sup>, among the most potent thus far described. 35O22 did not bind monomeric forms of Env tested, but did bind the trimeric BG505 SOSIP.664. Mutagenesis and a reconstruction by negative-stain electron microscopy of the Fab in complex with trimer revealed that it bound to a conserved epitope, which stretched across gp120 and gp41. The specificity of 35O22 represents a novel site of vulnerability on HIV Env, which serum analysis indicates to be commonly elicited by natural infection. Binding to this new site of vulnerability may thus be an important complement to current monoclonal-antibody-based approaches to immunotherapies, prophylaxis and vaccine design.**

Induction of a potent neutralizing antibody response capable of recognizing highly diverse isolates of HIV-1 is one of the most important goals of HIV vaccine research. This represents a considerable challenge given the extraordinary antigenic variability of the Env surface glycoprotein. However, approximately 20% of the HIV-infected population does develop a humoral immune response capable of recognizing highly diverse strains<sup>2–6</sup>. In the past several years improved patient cohorts<sup>2–6</sup>, HIV-specific B-cell isolation<sup>7–9</sup>, and IgG cloning techniques<sup>10,11</sup> have permitted extraordinary progress in the isolation of broadly neutralizing monoclonal antibodies (bNabs) from these individuals. Thus far, these primarily fall into four categories based on the position of their epitopes on the Env protein, a trimer of gp120 and gp41 heterodimers that is the target of neutralizing antibodies. These sites include the CD4-binding site on gp120 (refs 8, 12) (of which VRC01 is an example), the glycan-containing regions of V1V2 on gp120 (of which PG9 and PG16 are examples), the V3 region centred on the N332 glycan of gp120 (refs 7, 13) (of which PGT121 is an example) and the membrane-proximal external region (MPER) on gp41 (of which 10E8 is an example)<sup>14,15</sup>. It remains unclear to what extent these four categories represent the prevalent and immunodominant sites of Env vulnerability through which broad neutralizing responses are mediated, or whether additional specificities exist<sup>16–19</sup>.

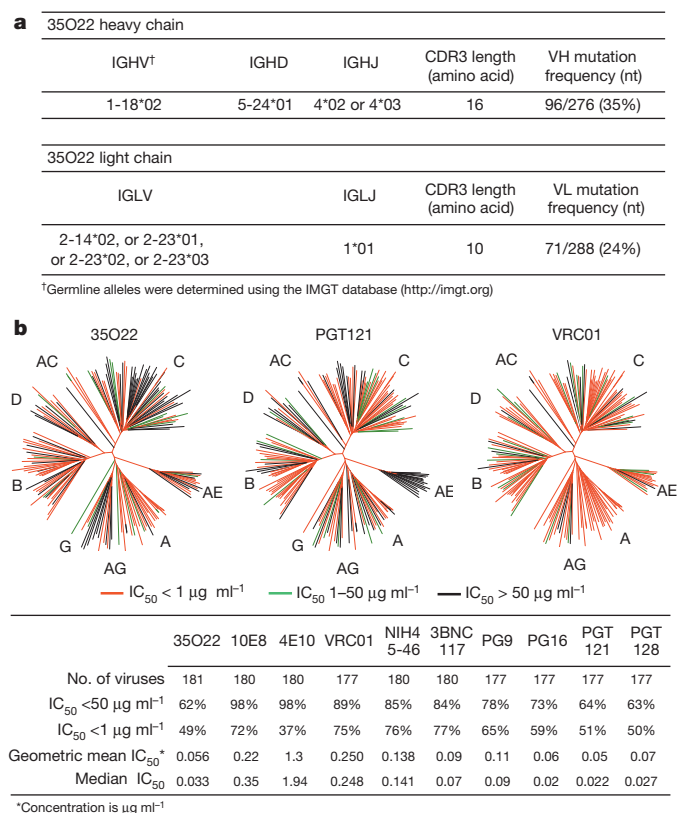
Here we report the isolation of a broad and potentially neutralizing HIV-specific monoclonal antibody, 35O22, that binds a novel epitope. The neutralizing activity of 35O22 is highly complementary to the activities of other known bNabs. We used mutagenesis, crystallography and electron

microscopy to define the Env site targeted by 35O22. Our results indicate that 35O22 neutralization occurs by a novel mode of trimer recognition along a conserved face on contiguous areas of gp41 and gp120.

For a better understanding of the specificities that underlie broadly neutralizing antibody responses we applied a technique to identify human monoclonal antibodies of interest from peripheral blood B cells without previous knowledge of the target specificity<sup>9</sup>. IgG<sup>+</sup> B cells of a donor (N152) with broad and potent neutralizing serum and from whom recently described 10E8 antibody was cloned<sup>20</sup> were sorted and expanded. The supernatants of B-cell microcultures were screened for neutralizing activity and IgG genes from positive wells were cloned and re-expressed. In addition to the 10E8 antibody, eight clonal family variants of an additional antibody with neutralization activity were found, among which the 35O22 antibody was the most potent and broad (Supplementary Table 1a, b). This antibody was derived from *IGHV-1-18\*02* and *IGLV-2-14\*02* germline genes, and was highly somatically mutated in variable genes of both heavy chain (35%) and  $\lambda$  light chain (24%) compared to germ line. The 35O22 antibody possessed a heavy-chain complementarity-determining 3 region (CDR H3) composed of 14 amino acids (Fig. 1a and Supplementary Table 2) and an insertion of 8 amino acids in framework 3 (FR3). High levels of somatic mutation and FR3 insertions are features of other HIV-specific bNabs<sup>7–9,12,13,21,22</sup>. Autoreactivity or poly-reactivity are properties of several HIV-specific antibodies<sup>23,24</sup> that could limit their use in therapies or prophylaxis. However, 35O22 bound HEP-2 epithelial cells only modestly (Extended Data Fig. 1a) and did not bind a panel of autoantigens (Extended Data Fig. 1b, c). Against a large panel of pseudoviruses, 35O22 neutralized 62% of 181 isolates with an IC<sub>50</sub> < 50 µg ml<sup>-1</sup> (Fig. 1b and Supplementary Table 3). In numerous cases where the IC<sub>50</sub> of 10E8 was > 1 µg ml<sup>-1</sup>, that of 35O22 was 100 to 1,000-fold lower (Supplementary Tables 1b and 3), indicating that their activities were highly complementary. It is likely that 35O22-like antibodies account for much of the breadth and potency of the N152 patient serum against clades A and B (Supplementary Table 3), whereas 10E8-like antibodies may account for much of the breadth against clade C isolates. Overall, the median IC<sub>50</sub> of 35O22 for sensitive viruses was 0.033 µg ml<sup>-1</sup>, which is among the most potent thus far described (Fig. 1b).

The neutralizing spectrum of 35O22 was then compared to those of other bNabs. The IC<sub>50</sub> of 35O22 against a panel of diverse isolates did not correlate with those of the bNabs VRC01, 10E8, PG9 and PGT121 (Extended Data Fig. 2a). In addition, a neutralization-based clustering analysis revealed that 35O22 clustered separately from other bNabs (Extended Data Fig. 2b). Furthermore, 35O22 did not compete with other known bNabs when bound to virus-like particles (VLPs) (Extended Data Fig. 2c). Its neutralization of many pseudoviruses did not exceed 80%

<sup>1</sup>Laboratory of Immunoregulation, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland 20892, USA. <sup>2</sup>Vaccine Research Center, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland 20892, USA. <sup>3</sup>The Scripps Center for HIV/AIDS Vaccine Immunology and Immunogen Discovery, The Scripps Research Institute, La Jolla, California 92037, USA. <sup>4</sup>International AIDS Vaccine Initiative (IAVI) Neutralizing Antibody Center, The Scripps Research Institute, La Jolla, California 92037, USA. <sup>5</sup>San Diego Biomedical Research Institute, San Diego, California 92121, USA. <sup>6</sup>Department of Medical Microbiology, Academic Medical Center, University of Amsterdam, Amsterdam 1100 DD, The Netherlands. <sup>7</sup>Department of Microbiology and Immunology, Weill Medical College of Cornell University, New York, New York 10065, USA. <sup>8</sup>Duke Human Vaccine Institute, Duke University, Durham, North Carolina 27710, USA.



**Figure 1 | Analyses of 35O22 sequence and neutralization.** **a**, Inferred germline genes encoding the variable regions of 35O22. **b**, Neutralizing activity of antibodies against a 181-isolate Env-pseudovirus panel. Dendrograms indicate the gp160 protein distance of HIV-1 primary isolate Env glycoproteins. Data below the dendrogram show the number of tested viruses, the percentage of viruses neutralized and the geometric mean or median IC<sub>50</sub> for viruses neutralized with an IC<sub>50</sub> < 50 µg ml<sup>-1</sup>.

even at high concentrations (Extended Data Fig. 3a) and its potency increased when pseudoviruses were produced in the presence of the glycosidase inhibitors NB-DNJ or kifunensine, consistent with recognition of high mannose (Extended Data Fig. 3b)<sup>25</sup>. However, neutralization was unaffected by mutation of *N*-linked glycosylation sites critical for binding of known bNabs (Extended Data Fig. 4a–c)<sup>7,13</sup>. Taken together, these data suggested that 35O22 binds glycans, but its specificity differed from all previously characterized bNabs.

Mutation of four predicted sites of *N*-linked glycosylation on HIV<sub>JRCSF</sub> (HIV-1/clade B) Env diminished neutralization potency—N88A, N230A, N241A and N625A (Fig. 2a and Supplementary Table 4). This result suggested that 35O22 recognized elements of both gp120 and gp41, a property that may be consistent with several recently isolated antibodies<sup>16–19</sup>. When mutations were introduced in the five residues on either side of these four sites, the V89A, T90A, K227A, T232A and S243A mutations each diminished neutralization (Supplementary Table 4). With the exception of V89A and K227A, it is likely that the impact of each of these mutations was to disrupt the Asn-X-Ser/Thr glycotransferase sequon. The T627A mutation had no impact, suggesting that a glycan may not be present at 625 and 35O22 may make a protein contact at this position. Overall, similar results were obtained using replication-competent HIV<sub>LAI</sub> (HIV-1/clade B) viruses (Supplementary Table 5). Examination of the sequences of resistant pseudoviruses within clade C did not reveal a clear pattern of variation at each of the positions found to affect 35O22 neutralization or within the glycosylation sequon. It is therefore possible that the resistance of clade C viruses is mediated by other factors such as variations in glycosylation pattern or conformation.

35O22 did not bind to a panel of soluble recombinant Env proteins (Extended Data Fig. 5a–c), suggesting that these do not have the appropriate

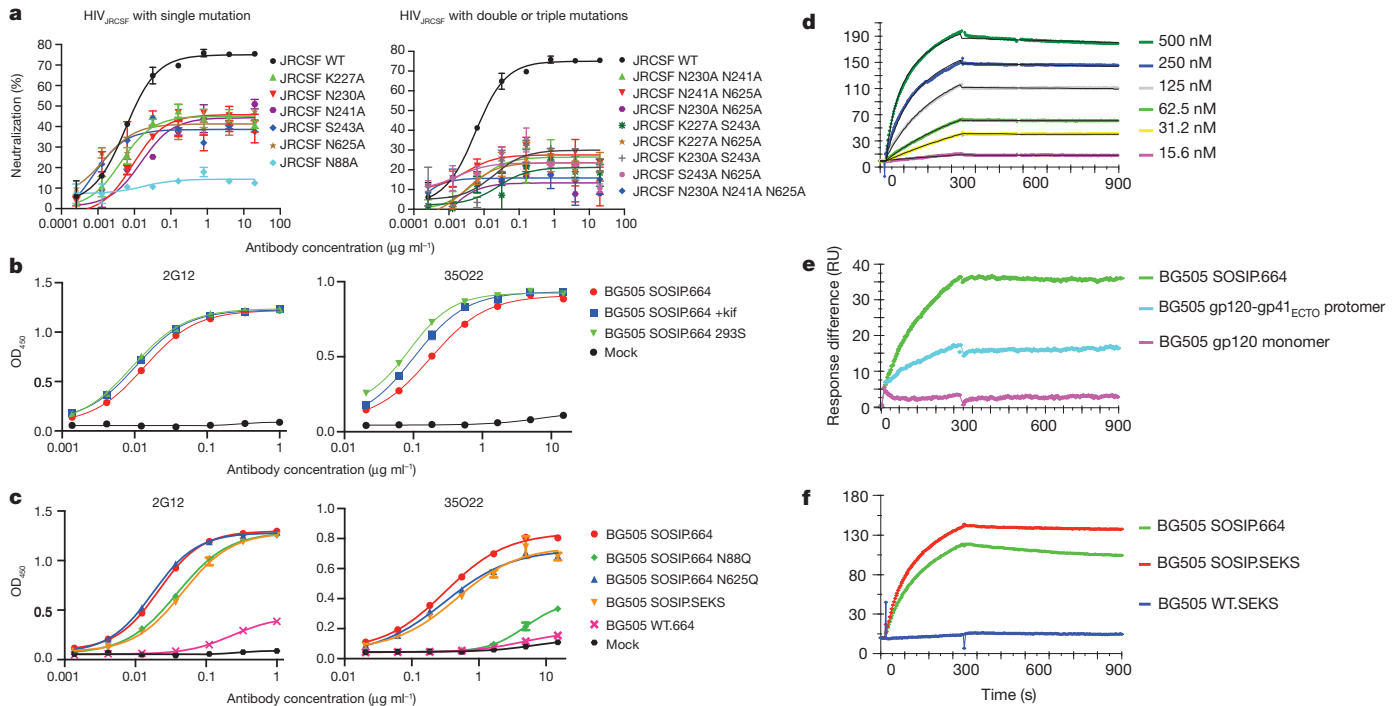
conformation or glycosylation for binding. However, the 35O22 antibody did bind a recently described stabilized, cleaved, soluble trimer, BG505 SOSIP.664 (Fig. 2b)<sup>26</sup>. Despite a plateau in neutralization below 50% (Extended Data Fig. 3a) and lacking glycans at positions 230 and 241, binding to BG505 SOSIP.664 trimer had numerous characteristics consistent with its activity against the HIV<sub>JRCSF</sub> pseudovirus. 35O22 binding was increased to trimer produced in cells treated with kifunensine or cells deficient in glycan processing (Fig. 2b), and diminished by mutations at positions 88 and 625 (Fig. 2c). 35O22 did bind to BG505 SOSIP trimer lacking the furin cleavage site (BG505 SOSIP.SEKS). In surface plasmon resonance (SPR) experiments, 35O22 also bound to immobilized BG505 SOSIP.664 with high affinity (dissociation constant ( $K_d$ ) = 5.6 nM) (Fig. 2d). Binding was markedly lower to the gp120–gp41<sub>ECTO</sub> protomer and no binding was detected to the gp120 monomer (Fig. 2e). 35O22 bound the uncleaved BG505 SOSIP.SEKS but no binding was observed to the uncleaved form lacking the SOSIP mutations (Fig. 2f). These observations, combined with the lack of binding of 35O22 to all other soluble forms of Env tested, suggested that this antibody requires a trimeric structure for binding its epitope on gp120 and gp41 (ref. 27).

To provide an atomic-level understanding of the structure of the 35O22 antibody, we crystallized the Fab of 35O22. Crystals were obtained that diffracted to 1.55 Å resolution (Supplementary Table 6). Overall the structure of 35O22 Fab revealed a relatively flat antigen-combining site, lacking long protruding loops, and flanked by the complementarity-determining region 1 on the light chain (CDR L1) and the 8-amino-acid insertion in FR3 of the heavy chain (Fig. 3a). The surface of the antigen-combining site was heavily altered by somatic mutation, and two pairs of cysteines introduced by somatic mutation in CDR L1 and L3 formed disulphide bonds (Fig. 3a and Extended Data Fig. 6a).

We next sought to determine the structure of the antibody–antigen complex. The ability of 35O22 to bind the BG505 SOSIP.664 trimer permitted imaging of the antibody–antigen complex by negative stain electron microscopy (EM). The reconstruction of these images showed that three 35O22 Fabs bound to the trimers at sites close to the predicted viral membrane (Fig. 3b and Extended Data Fig. 6b). Superposition of the negative stain reconstruction of the soluble BG505 SOSIP.664 with 35O22 Fab onto the BaL EM tomogram of the viral spike (Extended Data Fig. 6b) suggested that the viral membrane is in close contact to the 35O22 Fab light chain. Residues Tyr 68 and Trp 69 in the light chain and FR3 tyrosines at residues 65 and 72 form potential surfaces of membrane association. The 35O22 heavy chain was in close proximity to the four sites observed to contribute to the 35O22 epitope in mutagenesis experiments (N88, N230, N241 and N625). The CDR H3 was predicted to interact with N625 and CDR H2 with N88. The 8-amino-acid insertion in the framework 3 heavy chain is located close to residues 88–90 on gp120. Reversion of this insertion to germline markedly diminished neutralization against most pseudoviruses in our panel (Supplementary Table 7). 35O22 binds a surface on the Env spike that is distinct from two other antibodies, 8ANC195 and PGT151, reported to bind gp120 and gp41 (Extended Data Fig. 7)<sup>16–18</sup>.

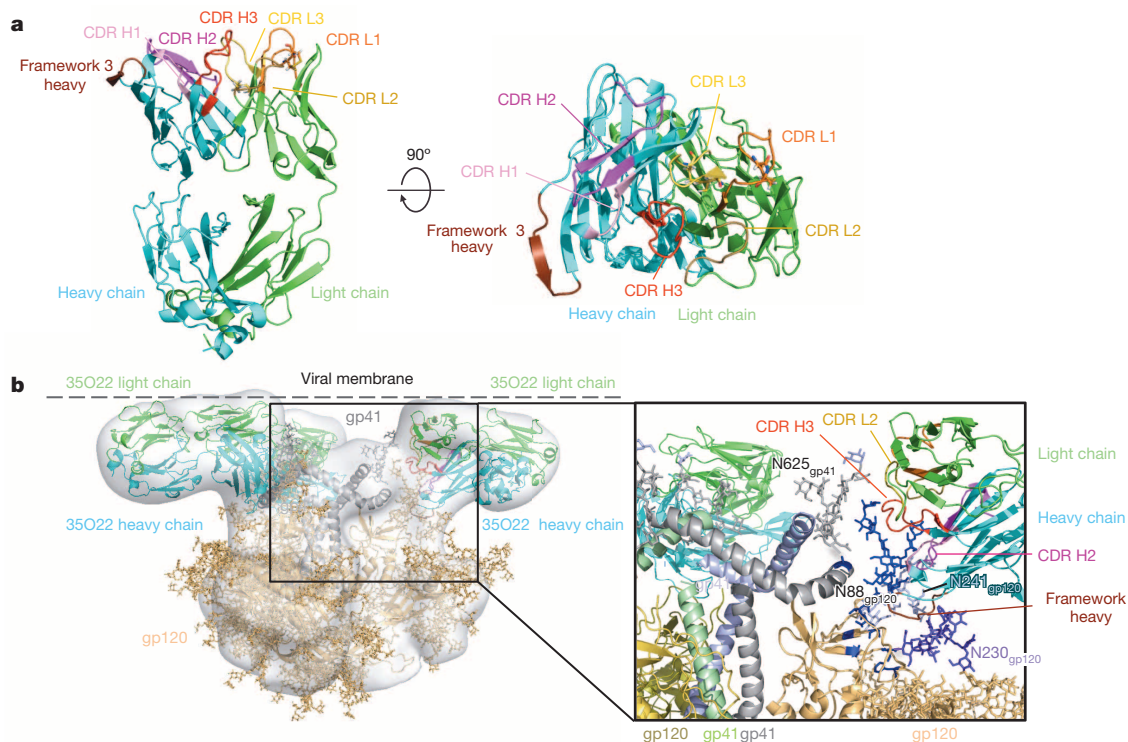
Analysis of the 35O22 site of vulnerability (Extended Data Fig. 8a–c) indicated that it is highly conserved. The glycans predicted at positions 88, 241 and 625 were found to be among the most highly conserved *N*-linked glycosylation sequons of 4,265 HIV-1 sequences in the Los Alamos database (Supplementary Table 8). Despite this high level of conservation, analysis of the Env gene of the patient's plasma virus showed that the predicted amino acid sequence varied at the critical 35O22 contacts. In addition to the previously published 10E8 escape mutations W680R and K683Q, an N230Q is predicted in one sequence, N241D in half of the sequences, and an N624D and N625Q in all sequences (Extended Data Fig. 9a). When these mutations were introduced into HIV<sub>JRCSF</sub> pseudoviruses, there was a drop in neutralization with the greatest effect caused by the N625Q mutation found in all of the plasma sequences (Extended Data Fig. 9b). These data suggest that the autologous virus has escaped neutralization by 35O22.





**Figure 2 | Binding specificity of 35O22.** **a**, Neutralization of HIV<sub>JRC5F</sub> pseudovirus or variants containing the indicated mutations. WT, wild type. **b**, Binding to BG505 SOSIP.664 trimer produced in cells treated with kifunensine or deficient in glycan processing (293S). OD<sub>450</sub>, optical density at 450 nm. **c**, Binding to BG505 trimers with the indicated mutations. BG505 SOSIP.SEKS lacks the furin cleavage site. BG505 WT.664 lacks stabilizing

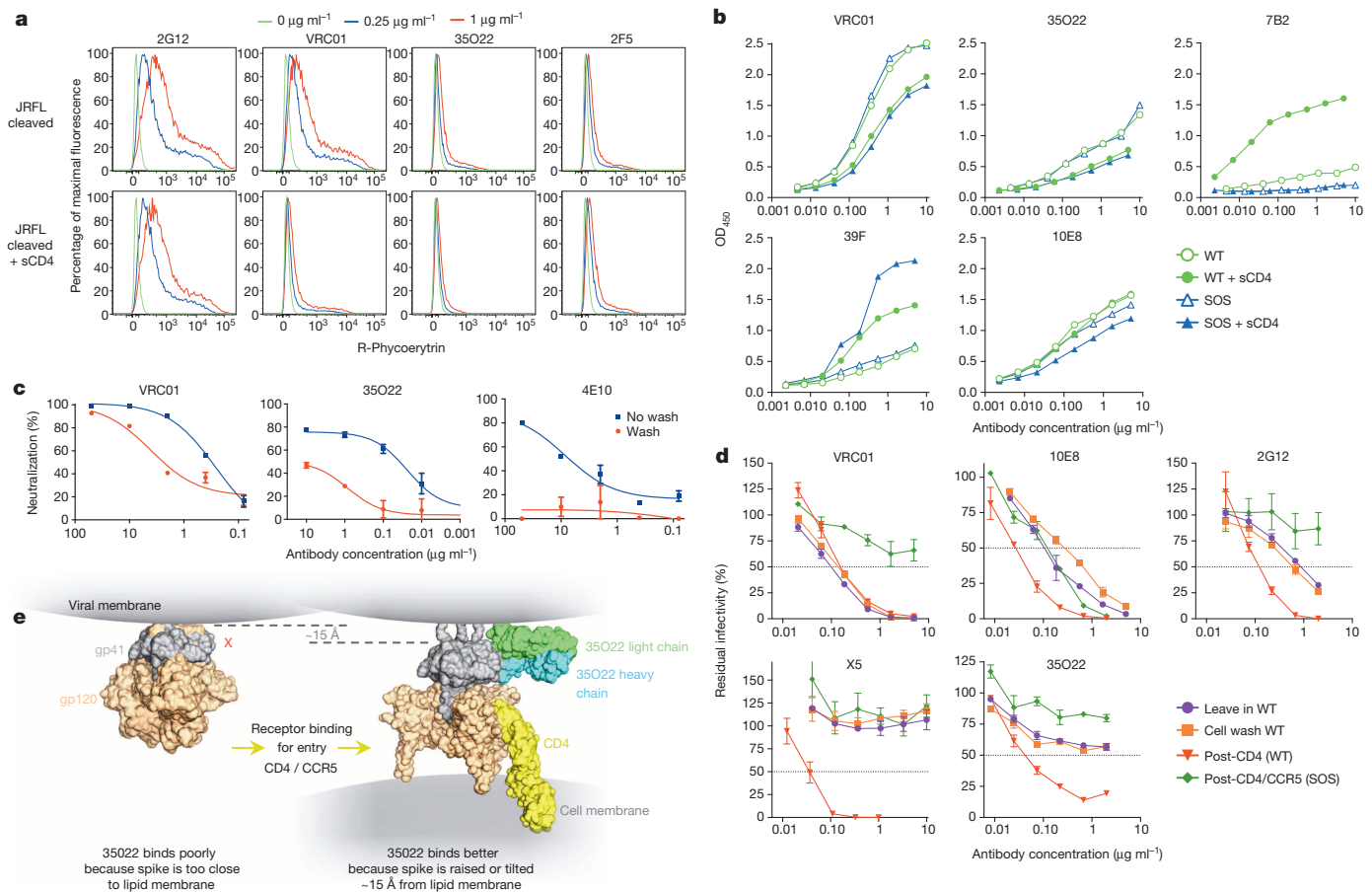
mutations and the antigen primarily represents gp41. **d**, SPR analysis of binding to immobilized BG505 SOSIP.664 trimers. **e**, Binding of 35O22 (250 nM) to BG505 SOSIP.664 trimers, gp120-gp41<sub>ECTO</sub> protomers, or monomeric gp120. **f**, Binding to BG505 SOSIP.664, BG505 SOSIP.SEKS, or the BG505 WT.SEKS lacking the SOSIP mutations. RU, response units.



**Figure 3 | Structure of Fab 35O22 and EM reconstruction in complex with HIV-1 Env.** **a**, Cartoon representation (left) of unbound 35O22 Fab. Heavy chain is in cyan, light chain in green, and the heavy chain FR3 insertion in brown. Right: 90° rotation, looking down on combining site. **b**, EM reconstruction of BG505 SOSIP.664 in complex with 35O22 Fab, with fitted

crystal structures. gp120 (light orange) and gp41 (grey) are shown with 35O22 (as in **a**). The approximate location of the viral membrane is indicated. Glycans N241 and N230 are not part of the BG505 sequence, but have been modelled for reference.





**Figure 4 | Binding or neutralization in the context of a lipid membrane.**

**a**, Staining of cell-surface-expressed HIV<sub>JRFL</sub> Env. **b**, ELISA assay of antibody binding to wild type or SOS HIV<sub>JRFL</sub> VLPs. The CD4-inducible 39F antibody or gp41-specific 7B2 are used as controls. **c**, Access to the HIV<sub>JRFL</sub> Env trimer on pseudovirions based on washing the antibody–pseudovirion mixture before

To gain insight into the prevalence of the specificity of 35O22, we added the 35O22 neutralization fingerprint to the ten that we had previously identified (Extended Data Fig. 8d, e)<sup>28</sup>. Notably, 13 of the sera (38%) showed significant 35O22 neutralization signals (>0.2). This level of prevalence was substantially higher than for the V1V2-directed response (typified by the PG9 antibody) or that of the 8ANC195 antibody. However, it was lower than the prevalence of responses to the MPER (50% prevalence), the CD4-binding site (53% prevalence), or the V3 glycan site (82% prevalence). The neutralizing activity of sera was also measured against HIV<sub>JRCSF</sub> pseudoviruses bearing N88A, N230A, N241A or N625A mutations (Extended Data Table 1). These mutations caused a greater than fivefold increase in ID<sub>50</sub> (50% inhibitory dilution) in more than half of donors, with a high level of concordance between the impact of each of these mutations within a given serum. These results suggested that 35O22 is unlikely to be the product of a unique B-cell repertoire or infecting virus, but rather arises commonly among patients that develop HIV-specific neutralizing antibodies.

To achieve a better understanding of the mechanism of 35O22's activity, we examined the timing of its binding and neutralization during virus fusion. Given the proximity of the 35O22 epitope to the membrane, it was important to perform these experiments in the context of Env expressed on cells or virions. In these settings, MPER-specific antibodies have limited access to the native trimer and bind best after the conformational changes induced by CD4 attachment<sup>29</sup>. Binding of 35O22 to Env expressed on the cell surface was low and similar to the MPER antibody 2F5 (Fig. 4a). 35O22 binding to VLPs was weak compared to that of VRC01 but similar to that of 10E8 (Fig. 4b). 35O22 binding was slightly inhibited by soluble CD4 (sCD4) binding, suggesting the 35O22

epitope is not induced by sCD4 under these experimental conditions. 35O22 neutralization was partially eliminated by washing of pseudovirions before infection, a result consistent with limited access to Env on free virions (Fig. 4c)<sup>29</sup>. However, if 35O22 was incubated with VLPs, permitted to bind to target cells, and then after 2 h the cells were washed, there was little impact on neutralization compared with the leave in format, similar to all other antibodies except the MPER 10E8 antibody (Fig. 4d). Similar to 2G12 and 10E8, 35O22 activity was relatively high in the post-CD4 format (see Methods), consistent with previous work showing that virus–sCD4 complexes are more sensitive to neutralization than virus alone<sup>30</sup>. In a post-CD4/CCR5 assay, only 10E8 neutralized virus, consistent with previous observations (Fig. 4d)<sup>30</sup>. Taken together with the structural data, these results suggest that in the context of a lipid membrane, 35O22 binds Env poorly before CD4 attachment. However, after trimer attachment to CD4, 35O22 may bind to an early intermediate that exposes the 35O22 epitope possibly by raising the Env spike within the viral membrane (Fig. 4e).

Our findings concerning the 35O22 antibody and its specificity have a number of implications for the use of antibodies in HIV therapy, prophylaxis and efforts to stimulate HIV-specific antibodies with vaccines. Its novel binding site and spectrum of activity against HIV strains suggest that it could complement other antibodies used in passive immunotherapy or prophylaxis. In addition, the antibody is extremely potent, indicating that its activity *in vivo* may therefore persist even at low concentrations. Perhaps most importantly, the novel epitope bound by the 35O22 antibody represents a new site of vulnerability that could potentially be targeted by HIV vaccines. The high prevalence of 35O22-like neutralizing activity in HIV-infected cohorts increases the likelihood

epitope is not induced by sCD4 under these experimental conditions. 35O22 neutralization was partially eliminated by washing of pseudovirions before infection, a result consistent with limited access to Env on free virions (Fig. 4c)<sup>29</sup>. However, if 35O22 was incubated with VLPs, permitted to bind to target cells, and then after 2 h the cells were washed, there was little impact on neutralization compared with the leave in format, similar to all other antibodies except the MPER 10E8 antibody (Fig. 4d). Similar to 2G12 and 10E8, 35O22 activity was relatively high in the post-CD4 format (see Methods), consistent with previous work showing that virus–sCD4 complexes are more sensitive to neutralization than virus alone<sup>30</sup>. In a post-CD4/CCR5 assay, only 10E8 neutralized virus, consistent with previous observations (Fig. 4d)<sup>30</sup>. Taken together with the structural data, these results suggest that in the context of a lipid membrane, 35O22 binds Env poorly before CD4 attachment. However, after trimer attachment to CD4, 35O22 may bind to an early intermediate that exposes the 35O22 epitope possibly by raising the Env spike within the viral membrane (Fig. 4e).

that production of similar antibodies could be induced by vaccination. In addition, the highly specific recognition by 35O22 of BG505 SOSIP.664 suggests that this soluble, cleaved trimer antigenically resembles the native Env trimer at the gp120–gp41 interface. Given the binding characteristics of 35O22, these results underscore the possibility that immunogens structurally similar to the native trimer are required for elicitation of such antibodies<sup>27</sup>.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 8 February; accepted 23 June 2014.**

**Published online 3 September; corrected online 5 November 2014 (see full-text HTML version for details).**

- Kwong, P. D. & Mascola, J. R. Human antibodies that neutralize HIV-1: identification, structures, and B cell ontogenies. *Immunity* **37**, 412–425 (2012).
- Doria-Rose, N. A. *et al.* Breadth of human immunodeficiency virus-specific neutralizing activity in sera: clustering analysis and association with clinical variables. *J. Virol.* **84**, 1631–1636 (2010).
- Sather, D. N. *et al.* Factors associated with the development of cross-reactive neutralizing antibodies during human immunodeficiency virus type 1 infection. *J. Virol.* **83**, 757–769 (2009).
- Walker, L. M. *et al.* A limited number of antibody specificities mediate broad and potent serum neutralization in selected HIV-1 infected individuals. *PLoS Pathog.* **6**, e1001028 (2010).
- Simek, M. D. *et al.* Human immunodeficiency virus type 1 elite neutralizers: individuals with broad and potent neutralizing activity identified by using a high-throughput neutralization assay together with an analytical selection algorithm. *J. Virol.* **83**, 7337–7348 (2009).
- Gray, E. S. *et al.* Antibody specificities associated with neutralization breadth in plasma from human immunodeficiency virus type 1 subtype C-infected blood donors. *J. Virol.* **83**, 8925–8937 (2009).
- Walker, L. M. *et al.* Broad and potent neutralizing antibodies from an African donor reveal a new HIV-1 vaccine target. *Science* **326**, 285–289 (2009).
- Wu, X. *et al.* Rational design of envelope identifies broadly neutralizing human monoclonal antibodies to HIV-1. *Science* **329**, 856–861 (2010).
- Huang, J. *et al.* Isolation of human monoclonal antibodies from peripheral blood B cells. *Nature Protocols* **8**, 1907–1915 (2013).
- Tiller, T. *et al.* Efficient generation of monoclonal antibodies from single human B cells by single cell RT-PCR and expression vector cloning. *J. Immunol. Methods* **329**, 112–124 (2008).
- Scheid, J. F. *et al.* Broad diversity of neutralizing antibodies isolated from memory B cells in HIV-infected individuals. *Nature* **458**, 636–640 (2009).
- Bonsignori, M. *et al.* Two distinct broadly neutralizing antibody specificities of different clonal lineages in a single HIV-1-infected donor: implications for vaccine design. *J. Virol.* **86**, 4688–4692 (2012).
- Walker, L. M. *et al.* Broad neutralization coverage of HIV by multiple highly potent antibodies. *Nature* **477**, 466–470 (2011).
- Burton, D. R. *et al.* Efficient neutralization of primary isolates of HIV-1 by a recombinant human monoclonal antibody. *Science* **266**, 1024–1027 (1994).
- Muster, T. *et al.* A conserved neutralizing epitope on gp41 of human immunodeficiency virus type 1. *J. Virol.* **67**, 6642–6647 (1993).
- Scharf, L. *et al.* Antibody 8ANC195 reveals a site of broad vulnerability on the HIV-1 envelope spike. *Cell Rep.* **7**, 785–795 (2014).
- Falkowska, E. *et al.* Broadly neutralizing HIV antibodies define a glycan-dependent epitope on the prefusion conformation of gp41 on cleaved envelope trimers. *Immunity* **40**, 657–668 (2014).
- Blattner, C. *et al.* Structural delineation of a quaternary, cleavage-dependent epitope at the gp41–gp120 interface on intact HIV-1 Env trimers. *Immunity* **40**, 669–680 (2014).
- Zhang, M. Y. *et al.* Identification and characterization of a broadly cross-reactive HIV-1 human monoclonal antibody that binds to both gp120 and gp41. *PLoS ONE* **7**, e44241 (2012).
- Huang, J. *et al.* Broad and potent neutralization of HIV-1 by a gp41-specific human antibody. *Nature* **491**, 406–412 (2012).
- Scheid, J. F. *et al.* Sequence and structural convergence of broad and potent HIV antibodies that mimic CD4 binding. *Science* **333**, 1633–1637 (2011).
- Klein, F. *et al.* Somatic mutations of the immunoglobulin framework are generally required for broad and potent HIV-1 neutralization. *Cell* **153**, 126–138 (2013).
- Haynes, B. F. *et al.* Cardiophilic polyspecific autoreactivity in two broadly neutralizing HIV-1 antibodies. *Science* **308**, 1906–1908 (2005).
- Mouquet, H. *et al.* Polyreactivity increases the apparent affinity of anti-HIV antibodies by heterologation. *Nature* **467**, 591–595 (2010).
- Doores, K. J. & Burton, D. R. Variable loop glycan dependency of the broad and potent HIV-1-neutralizing antibodies PG9 and PG16. *J. Virol.* **84**, 10510–10521 (2010).
- Sanders, R. W. *et al.* A next-generation cleaved, soluble HIV-1 Env Trimer, BG505 SOSIP.664 gp140, expresses multiple epitopes for broadly neutralizing but not non-neutralizing antibodies. *PLoS Pathog.* **9**, e1003618 (2013).
- Ringe, R. P. *et al.* Cleavage strongly influences whether soluble HIV-1 envelope glycoprotein trimers adopt a native-like conformation. *Proc. Natl Acad. Sci. USA* **110**, 18256–18261 (2013).
- Georgiev, I. S. *et al.* Delineating antibody recognition in polyclonal sera from patterns of HIV-1 isolate neutralization. *Science* **340**, 751–756 (2013).
- Pancera, M. & Wyatt, R. Selective recognition of oligomeric HIV-1 primary isolate envelope glycoproteins by potentially neutralizing ligands requires efficient precursor cleavage. *Virology* **332**, 145–156 (2005).
- Crooks, E. T. *et al.* Characterizing anti-HIV monoclonal antibodies and immune sera by defining the mechanism of neutralization. *Hum. Antibodies* **14**, 101–113 (2005).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** We thank J. P. Moore for experiments using the BG505 SOSIP trimer, discussions of experimental data, and review of this manuscript. We thank K. Lloyd, R. Parks, J. Eudailey and J. Blinn for performing autoantibody assays. We also thank S. Moir for providing patient samples. This project has been funded in part with federal funds from the Intramural Research Programs of NIAID, National Institutes of Health. This work was also supported by the NIH HIVRAD grant P01 AI082362, and by the Aids Fonds Netherlands, grants 2011032 and 2012041. R.W.S. is a recipient of a Vidi grant from the Netherlands Organization for Scientific Research (NWO) and a Starting Investigator Grant from the European Research Council (ERC-StG-2011-280829-SHEV). J.H.L., Y.F., R.T.W. and A.W. are supported through the International AIDS Vaccine Initiative. J.M.B. is supported by NIH grants AI93278 and AI84714. Use of sector 22 (Southeast Region Collaborative Access team) at the Advanced Photon Source was supported by the US Department of Energy, Basic Energy Sciences, Office of Science, under contract number W-31-109-Eng-38. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government.

**Author Contributions** M.C., J.H., B.H.K., M.P., J.R.M., P.D.K., J.M.B., R.T.W., R.W.S. and A.B.W. each contributed to the design of the study, analysis of the data, and preparation of this manuscript. J.H. and B.H.K. performed B-cell sorting, antibody cloning, neutralization assays, cell surface staining and epitope mapping assays. H.L. performed the sequencing and analysis of the patient plasma virus. L.L. cultured B cells and assisted with recovery of IgG genes. N.A.D.-R. performed the 35O22 biotinylation and provided data of PG9, PG16, PGT121 and PGT128 breadth and potency. P.-J.K. performed SPR experiments. R.D., K.S., A.T.d.I.P. and P.P. performed the ELISA BG505 SOSIP.664 mutants. M.J.v.G. performed neutralization assays with HIV<sub>LAI</sub> viruses. M.A. and B.F.H. performed the autoreactivity assays. I.S.G., A.D. and G.-Y.C. performed serological analysis. M.P. and P.D.K. performed 35O22 crystallographic analysis. J.H.L. and A.B.W. performed the negative stain EM studies. J.M.B. and T.T. planned and performed the antibody competition, and HIV VLP entry studies. Y.F. and R.T.W. planned and performed the washout and kinetic studies. S.A.M. led the clinical care of the patients. R.T.B. screened the B-cell culture supernatants for neutralization activity.

**Author Information** The 35O22 heavy and light chain plasmids and antibody have been submitted to the NIH-AIDS reagent program. The nucleotide sequence of 35O22 and its variants have been submitted to GenBank under accession numbers KM001872–KM001879 (heavy chain) and KM001880–KM001887 (light chain). The Env nucleotide sequences of the patient (N152) plasma virus are deposited in GenBank under accession numbers KM516886–KM516897. Coordinates and structure factors for 35O22 Fab have been deposited with the Protein Data Bank under accession code 4TOY. The reconstruction of BG505 SOSIP.664 in complex with 35O22 Fab has been deposited in the Electron Microscopy Data Base under accession code EMD-2672. Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.C. ([mconnors@nih.gov](mailto:mconnors@nih.gov)).

## METHODS

**Study patients.** We selected the plasma and peripheral blood mononuclear cells (PBMCs) from the HIV-1-infected patients enrolled in the National Institute of Health under a clinical protocol (ClinicalTrials.gov; <https://clinicaltrials.gov/ct2/show/NCT00029445?term=NCT00029445&rank=1> identifier NCT00029445) approved by the Investigational Review Board in the National Institute of Allergy and Infectious Diseases (NIAID-IRB). All participants signed informed consent approved by the NIAID-IRB. The criteria for enrolment were as follows: having a detectable viral load, a stable CD4 T-cell count above 400 cells  $\mu\text{L}^{-1}$ , being diagnosed with HIV infection for at least 4 years, and off antiretroviral treatment for at least 5 years. Donor N152 was selected for B-cell sorting and antibody generation because his serum neutralizing activity is among the most potent and broad in our cohort. At the time of leukapheresis, he had been infected with HIV-1 for 20 years, with CD4 T-cell counts of 325 cells  $\mu\text{L}^{-1}$ , plasma HIV-1 RNA values of 3,811 copies  $\text{mL}^{-1}$  and was not on antiretroviral treatment.

**Memory B-cell staining, sorting and antibody cloning.** Staining and single-cell sorting of memory B cells were performed following a detailed protocol recently published<sup>9</sup>. Briefly, a total of 140,000 CD19<sup>+</sup>IgA<sup>−</sup>IgD<sup>−</sup>IgM<sup>−</sup> memory B cells were sorted and re-suspended in medium with IL-2, IL-21 and irradiated 3T3-msCD40L feeder cells, and seeded into 384-well microtitre plates at a density of 4 cells per well. After 13 days of incubation, supernatants from each well were screened for neutralization activity using a high-throughput micro-neutralization assay against HIV-1<sub>MN.3</sub> and HIV-1<sub>BaL.26</sub>. From the wells that scored positive in both the HIV-1<sub>MN.3</sub> and HIV-1<sub>BaL.26</sub> neutralization assay the variable region of the heavy chain and the light chain of the immunoglobulin gene were amplified by RT-PCR and re-expressed as described previously<sup>10</sup>. The full-length IgG was purified using a recombinant protein-A column (GE Healthcare).

**Generation of pseudoviruses and VLPs.** HIV-1 Env pseudoviruses were generated by co-transfection of 293T cells with pSG3 delta Env backbone and a second plasmid that expressed HIV-1 Env at a ratio of 2:1. HIV<sub>LA1</sub> viruses were generated by transfection of a single plasmid containing the entire viral genome. 72 h after transfection, supernatants containing pseudoviruses were harvested and frozen at  $-80^{\circ}\text{C}$  until further use. Glycosidase inhibitors were added to the cells at the time of transfection at flowing concentration, 25  $\mu\text{M}$  kifunensine, 500  $\mu\text{M}$  NB-DNJ and 20  $\mu\text{M}$  swainsonine. JRCSF mutants were produced by altering the JRCSF Env plasmid using QuikChange Lightning mutagenesis according to the manufacturer's protocol (Agilent). VLPs for ELISA assays were produced by transient transfection of 293T cells with a pCAGGS Env-expressing plasmid and the subgenomic plasmid pNL4-3.Luc.R-E- as previously described<sup>31</sup>. The N152 patient autologous virus was sequenced as previously described<sup>20</sup>.

**Neutralization assays.** Neutralization activity of monoclonal antibodies or serum was measured using single-round HIV-1 Env-pseudovirus infection of TZM-bl cells as described previously<sup>32</sup>. Heat-inactivated patient serum or monoclonal antibody was serially diluted fivefold with Dulbecco's modified Eagle medium–10% FBS (Gibco), and 10  $\mu\text{L}$  was incubated with 40  $\mu\text{L}$  of pseudovirus in a 96-well plate at  $37^{\circ}\text{C}$  for 30 min. TZM-bl cells were then added and plates were incubated for 48 h. Assays were then developed with a luciferase assay system (Promega), and the relative light units (RLU) were read on a luminometer (Perkin Elmer). Washing of pseudovirions to determine access by antibody to the trimer on free virus was performed as described previously<sup>33</sup>. Neutralization in various formats to determine the timing and mechanism of neutralization were performed as described previously<sup>30,34</sup>. Briefly, in the standard 'leave in' format monoclonal antibodies were mixed with VLPs 1 h before infection of CF2.CD4.CCR5 cells with no subsequent washing. Alternatively, in a 'cell wash' format, virus and antibody were incubated with target cells for 2 h, followed by a wash to remove any unattached virus. In the post-CD4 format, VLPs were premixed with 3  $\mu\text{g mL}^{-1}$  of sCD4 for 15 min, then incubated with antibody for 1 h before adding to cells that lack CD4 but express CCR5 (CF2. syn.CCR5 cells). In the post-CD4/CCR5 format, SOS VLPs were permitted to attach to cells for 2 h, unbound VLPs were washed away and graded concentrations of antibodies were added before infection was activated with DTT.

**Prediction of prevalence of 35O22-like antibodies in patient serum.** An antibody neutralization fingerprint, the pattern with which an antibody neutralizes a panel of diverse viral strains, was used to delineate the structural epitope recognized by that antibody<sup>28,35</sup>. For each serum, neutralization on a panel of 21 HIV-1 strains was represented as a combination of the neutralization fingerprints for a reference set of antibody specificities targeting the other four major sites of Env vulnerability as well as the 35O22 fingerprint, to obtain an estimate of the relative contribution of each of these specificities to neutralization by the given serum. The neutralization behaviour of sera was deconvoluted from a cohort of 34 donors, each with neutralization breadths of greater than 50%.

**Binding assays.** HIV<sub>YU2</sub> gp160 extracellular domain (gp140) foldon trimer, gp120 and gp41 monomers were produced as described previously<sup>36</sup>. HIV<sub>BaL</sub> (clade B), HIV<sub>CM235</sub> (CRF01\_AE), HIV<sub>CN54</sub> (clade C), HIV<sub>96ZM651</sub> (clade C), HIV<sub>93TH975</sub>

(clade E) gp120 proteins, and the HIV<sub>UG37</sub> (clade A), HIV<sub>CN54</sub> (clade C), HIV<sub>UG21</sub> (clade D) and HIV<sub>BR29</sub> (clade F) gp140 monomers were obtained through the NIH AIDS Research and Reagent Program. For ELISA assays each antigen ( $2 \mu\text{g mL}^{-1}$ ) was coated on 96-well plates overnight at  $4^{\circ}\text{C}$ . Plates were blocked with BLOTTO buffer (PBS, 1% FBS, 5% non-fat milk) for 1 h at room temperature, followed by incubation with antibody serially diluted in disruption buffer (PBS, 5% FBS, 2% BSA, 1% Tween-20) for 1 h at room temperature. 1:10,000 dilution of horseradish peroxidase (HRP)-conjugated goat anti-human IgG antibody was added for 1 h at room temperature. Plates were washed between each step with 0.2% Tween 20 in PBS. Plates were developed using 3,3',5,5'-tetramethylbenzidine (TMB) (Sigma) and read at 450 nm. ELISA assays using the BG505 SOSIP.664 trimer and mutants were performed as described previously<sup>26,27</sup>. Surface plasmon resonance experiments were performed at  $25^{\circ}\text{C}$  on a Biacore 3000 instrument (GE Healthcare) using D7324 or anti-histidine capture as previously described<sup>37</sup>. For the SPR experiment presented in Fig. 2d an anti-histidine capture was used (amount of immobilized ligand ( $R_L$ ) =  $\sim 500$  RU) and in Fig. 2e, f a D7324 capture was used ( $R_L$  =  $\sim 150$  RU for trimer and protomer; 130 RU for gp120 in Fig. 2e and  $R_L$  =  $\sim 500$  RU in Fig. 2f). ELISA assays using VLPs were performed as previously described<sup>30,31</sup>.

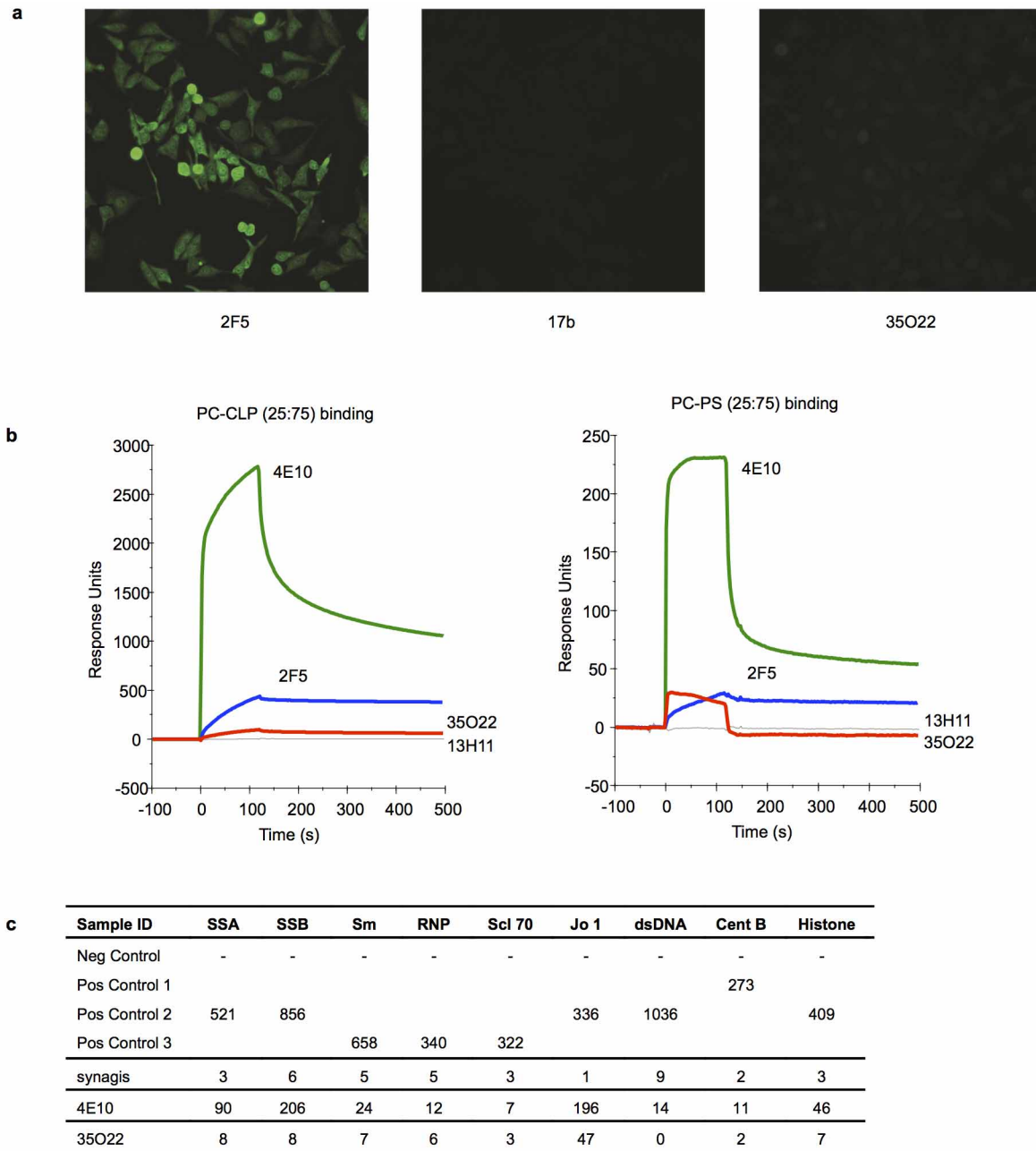
**Autoreactivity assays.** Reactivity to HIV-1 negative human epithelial (HEP-2) cells was determined by indirect immunofluorescence on slides using Evans Blue as a counterstain and FITC-conjugated goat anti-human IgG (Zeus Scientific)<sup>23</sup>. Slides were photographed on a Nikon Optiphot fluorescence microscope. Kodachrome slides were taken of each monoclonal antibody binding to HEP-2 cells at a 10-s exposure, and the slides scanned into digital format. The Luminex AteNA Multi-Lyte ANA test (Wampole Laboratories) was used to test for monoclonal antibody reactivity to SSA/Ro, SS-B/La, Sm, ribonucleoprotein (RNP), Jo-1, double-stranded DNA, centromere B, and histone and was performed as per the manufacturer's specifications and as previously described<sup>23</sup>. Monoclonal antibody concentrations assayed were  $50 \mu\text{g mL}^{-1}$ . 10  $\mu\text{L}$  of each concentration were incubated with the luminex fluorescent beads and the test performed per the manufacturer's specifications.

**Structure determination and analysis.** The antigen-binding fragment of 35O22 (Fab) was prepared using HRV3C digestion, as previously described<sup>38</sup>. HRV3C was introduced in the hinge region of the heavy chain plasmid DNA. Both light and heavy chain plasmids were co-transfected in 293F as described previously. The antibody, 35O22 with HRV3C IgG was purified over protein A, cleaved with HRV3C and the flow-through collected and run onto a size-exclusion chromatography (S200). Purified 35O22 Fab set up for  $20^{\circ}\text{C}$  vapour diffusion sitting-drop crystallizations on the Honeybee 963 robot. A total of 576 initial conditions adapted from the commercially available Hampton (Hampton Research), Precipitant Synergy (Emerald Biosystems) and Wizard (Emerald Biosystems) crystallization screens were set up and imaged using the Rockimager (Formulatrix), followed by hand optimization of crystal hits. Crystals were grown in 15% isopropanol, 25% PEG 3350, 0.2 M ammonium citrate pH 4.5 diffracted to 1.55 Å resolution in a cryoprotectant composed of mother liquor supplemented with 15% 2R-3R-butenediol. After mounting the crystals on a loop, they were flash cooled and data were collected at 1.00 Å wavelength at SERCAT ID-22 beamlines (APS) and processed using HKL-2000<sup>39</sup>. Structures were solved through molecular replacement with Phaser<sup>40,41</sup>, using a previously obtained free structure of VRC01 germ line as a search model. Structure solution identified one Fab per asymmetric unit in a  $P4_12_1$  lattice. Refinement of the structure was undertaken with Phenix, with iterative model building using Coot<sup>42</sup>. Refinement resulted in an  $R_{\text{cryst}}$  value of 16.65% ( $R_{\text{free}}$  = 18.22%). The structure was validated with MolProbity<sup>43</sup>, yielding 98.1% and 100% of residues falling within most favoured Ramachandran regions and allowed Ramachandran regions, respectively. All graphics were prepared with Pymol (PyMOL Molecular Graphics System).

**Electron microscopy and image processing.** Negative stain EM grids were prepared as previously stated<sup>44</sup>. Data were collected using a FEI Tecnai T12 electron microscope operating at 120 keV, with an electron dose of  $\sim 30 \text{ e}^{-} \text{Å}^{-2}$  and a magnification of 52,000 $\times$  that resulted in a pixel size of 2.05 Å at the specimen plane. Images were acquired with a Tietz TemCam-F416 CMOS camera using a nominal defocus range of 800 to 1,000 nm using the Legikon interface. Image processing was carried out as described previously<sup>45</sup>. The final reconstruction was performed using 4,746 unbinned particles, refining for 40 iterations with C3 symmetry applied. The resulting density was  $\sim 19$  Å resolution at a Fourier shell correlation (FSC) cutoff of 0.5. Fab fitting was carried out using the Fit function in Chimera, using 0.0115 contour level for the map. The Fab orientation where the heavy chain is towards gp120 (as shown in Fig. 3b) was chosen on the basis of the two correlation coefficients: 0.95 (471 of 7,058 atoms outside contour) versus 0.91 (1,012 of 7,058 atoms outside contour). In the alternate orientation (light chain towards gp120), the FR3 insertion did not fit within the density. The man-9 glycans were modelled onto the BG505 SOSIP.R6.664 (PDB ID 4NCO)<sup>46</sup> using the GLYCAM-Web server (<http://www.glycam.com>). Dominant sites of vulnerability to neutralizing antibody elicited during chronic infection in Extended Data Fig. 7 is shown in the context of an EM tomogram from the HIV<sub>BaL</sub> viral spike<sup>47</sup>.

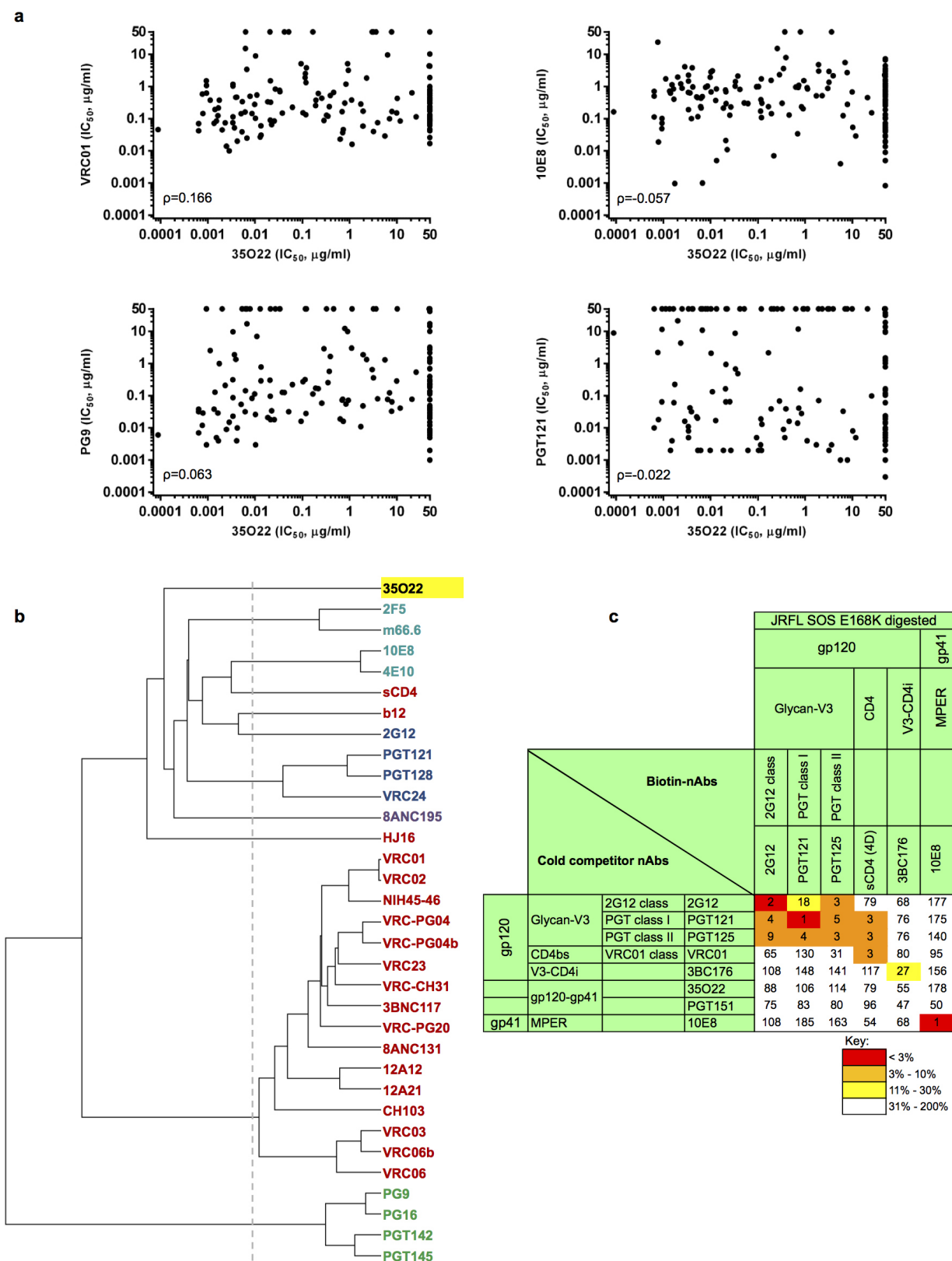
31. Tong, T., Crooks, E. T., Osawa, K. & Binley, J. M. HIV-1 virus-like particles bearing pure env trimers expose neutralizing epitopes but occlude nonneutralizing epitopes. *J. Virol.* **86**, 3574–3587 (2012).
32. Li, M. *et al.* Human immunodeficiency virus type 1 env clones from acute and early subtype B infections for standardized assessments of vaccine-elicited neutralizing antibodies. *J. Virol.* **79**, 10108–10125 (2005).
33. Chakrabarti, B. K. *et al.* Direct antibody access to the HIV-1 membrane-proximal external region positively correlates with neutralization sensitivity. *J. Virol.* **85**, 8217–8226 (2011).
34. Binley, J. M. *et al.* Redox-triggered infection by disulfide-shackled human immunodeficiency virus type 1 pseudovirions. *J. Virol.* **77**, 5678–5684 (2003).
35. Chuang, G. Y. *et al.* Residue-level prediction of HIV-1 antibody epitopes based on neutralization of diverse viral strains. *J. Virol.* **87**, 10047–10058 (2013).
36. Pancera, M. *et al.* Soluble mimetics of human immunodeficiency virus type 1 viral spikes produced by replacement of the native trimerization domain with a heterologous trimerization motif: characterization and ligand binding analysis. *J. Virol.* **79**, 9954–9969 (2005).
37. Yasmeen, A. *et al.* Differential binding of neutralizing and non-neutralizing antibodies to native-like soluble HIV-1 Env trimers, uncleaved Env proteins, and monomeric subunits. *Retrovirology* (in the press) (2014).
38. McLellan, J. S. *et al.* Structure of HIV-1 gp120 V1/V2 domain with broadly neutralizing antibody PG9. *Nature* **480**, 336–343 (2011).
39. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Macromol. Crystallogr. A* **276**, 307–326 (1997).
40. McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).
41. Winn, M. D. *et al.* Overview of the CCP4 suite and current developments. *Acta Crystallogr. D* **67**, 235–242 (2011).
42. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60**, 2126–2132 (2004).
43. Davis, I. W. *et al.* MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res.* **35**, W375–W383 (2007).
44. Kong, L. *et al.* Supersite of immune vulnerability on the glycosylated face of HIV-1 envelope glycoprotein gp120. *Nature Struct. Mol. Biol.* **20**, 796–803 (2013).
45. Thornburg, N. J. *et al.* Human antibodies that neutralize respiratory droplet transmissible H5N1 influenza viruses. *J. Clin. Invest.* **123**, 4405–4409 (2013).
46. Julien, J. P. *et al.* Crystal structure of a soluble cleaved HIV-1 envelope trimer. *Science* **342**, 1477–1483 (2013).
47. Liu, J., Bartesaghi, A., Borgnia, M. J., Sapiro, G. & Subramaniam, S. Molecular architecture of native HIV-1 gp120 trimers. *Nature* **455**, 109–113 (2008).





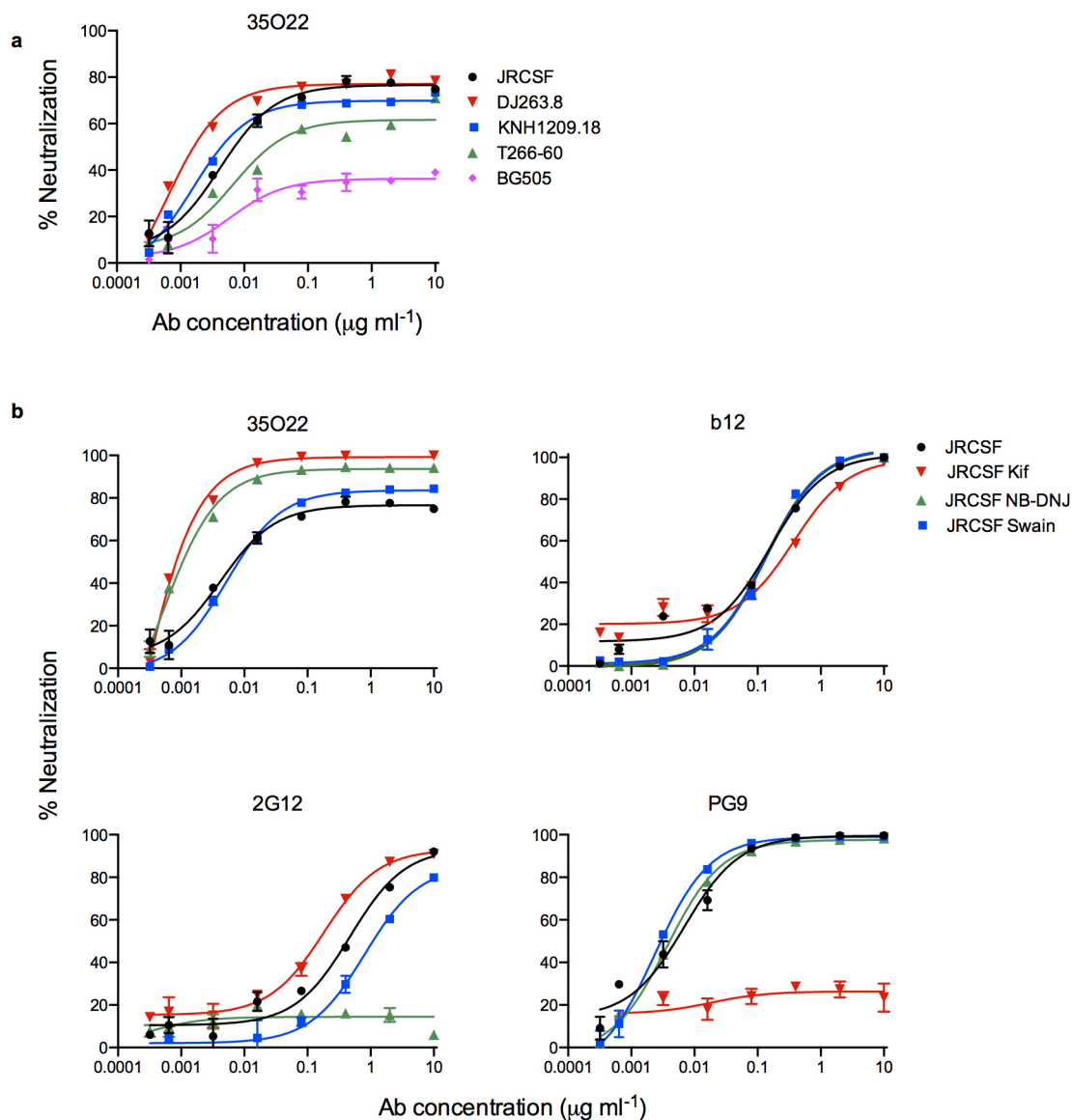
**Extended Data Figure 1 | Analysis of 35O22 autoreactivity.** **a**, Reactivity of 35O22 with HEP-2 epithelial cells. 2F5 was used as a positive control and 17b as a negative control. Antibody concentration was  $25 \mu\text{g ml}^{-1}$ . All pictures are shown at  $400\times$  magnification. **b**, SPR analysis of 35O22 binding to anionic phospholipids. 35O22 was injected over PC-CLP liposomes or PC-PS liposomes immobilized on the BIAcore L1 sensor chip. 4E10 and 2F5 were used

as positive controls and 13H1 as a negative control. **c**, Reactivity of 35O22 with autoantigens detected in Luminex assay. 4E10 was used as a positive control. Synagis, an anti-RSV monoclonal antibody, was used as a negative control. SSA, Sjogren's syndrome antigen A; SSB, Sjogren's syndrome antigen B; Sm, Smith antigen; RNP, ribonucleoprotein; Scl 70, scleroderma 70; Jo1, antigen; CentrB, centromere B. A positive response is  $>120$  units.



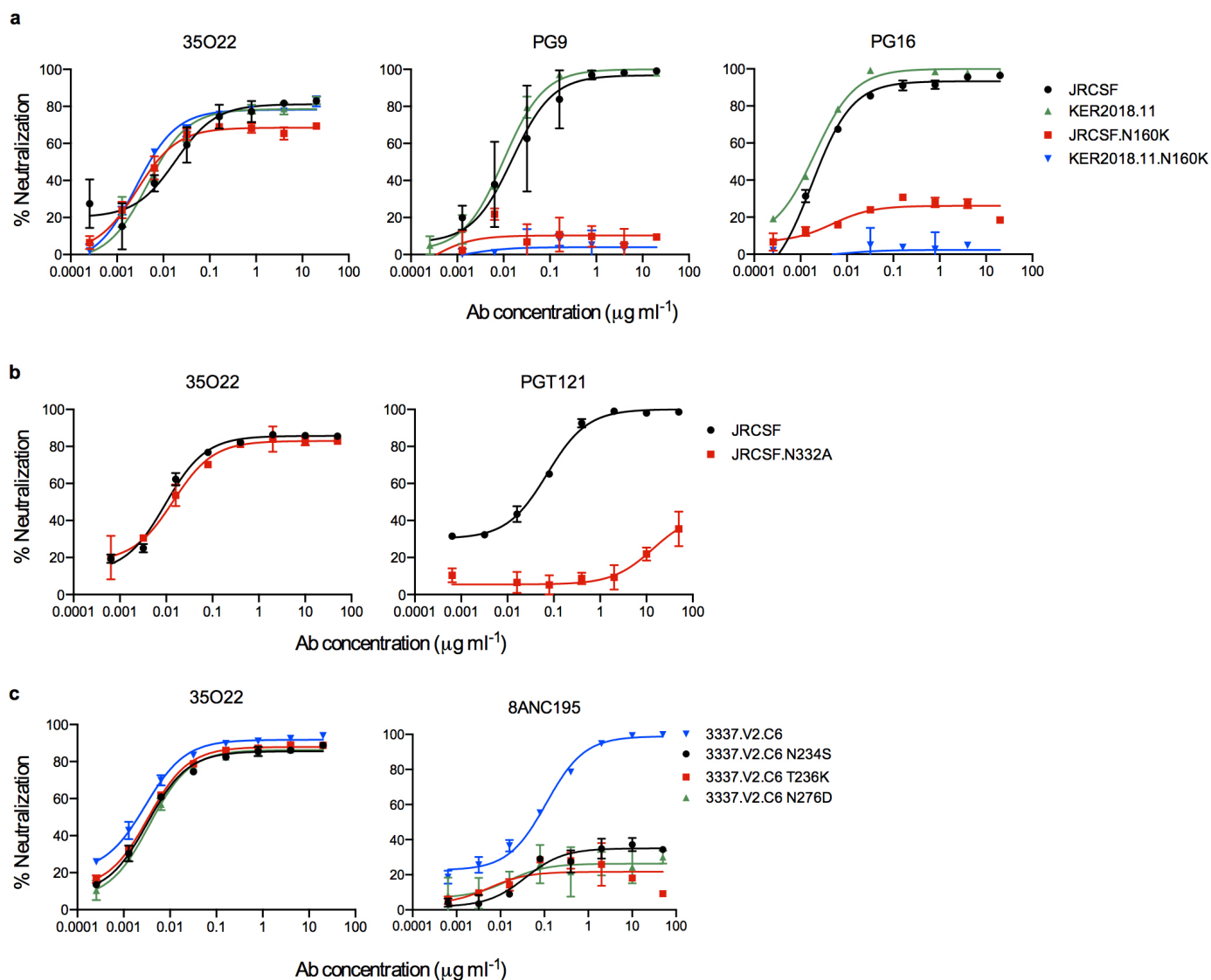
**Extended Data Figure 2 | Neutralization similarities between 35O22 and other HIV-1 bNAbs.** **a**, Correlation (Spearman) between the neutralization potencies of 35O22 and the indicated antibody against 172 pseudoviruses. Representatives from all four major sites of vulnerability are shown. Resistant strains corresponding to values of  $>50 \mu\text{g ml}^{-1}$  are plotted as 50. **b**, Neutralization-based clustering of bNAbs over a set of 172 diverse HIV-1 strains. A putative epitope-specific clustering cutoff is shown as a dashed line. Antibodies are coloured according to the respective target site of vulnerability:

red, CD4bs; blue, glycan-V3; green, V1V2; light blue, MPER; purple, other. 35O22 (yellow) clusters separately from all other antibodies, indicating a novel mechanism of neutralization. **c**, 35O22 competition with other bNAbs on HIV<sub>JRFL</sub> VLPs with the trimer stabilizing SOS mutations in an ELISA assay. Biotin-bNAbs were titrated into the ELISA at increasing concentrations in the presence of excess ( $10 \mu\text{g ml}^{-1}$ ) cold competitor neutralizing antibodies. Values in the table indicate percentage binding of biotin-nAbs in the presence of cold competitor.



**Extended Data Figure 3** | 35O22 binds to *N*-linked glycans. **a**, Neutralization by 35O22 plateaus below 80% against several pseudoviruses. **b**, Neutralization activity of monoclonal antibodies against JRCSF pseudoviruses generated in

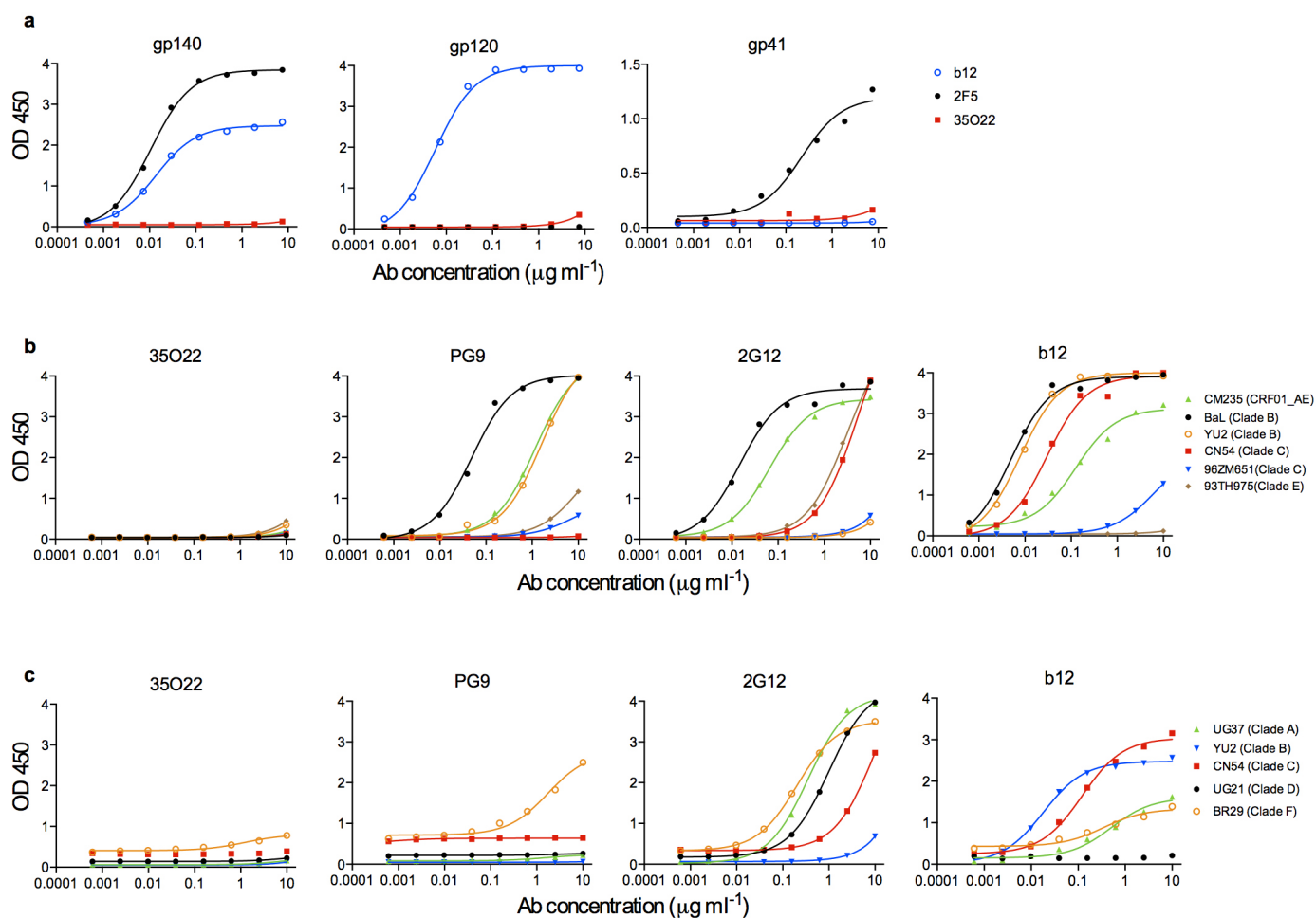
the presence of glycosidase inhibitors, such as kifunensine (25  $\mu\text{M}$ ), NB-DNJ (500  $\mu\text{M}$ ) or swainsonine (20  $\mu\text{M}$ ). Error bars denote one standard error of the mean (s.e.m.).



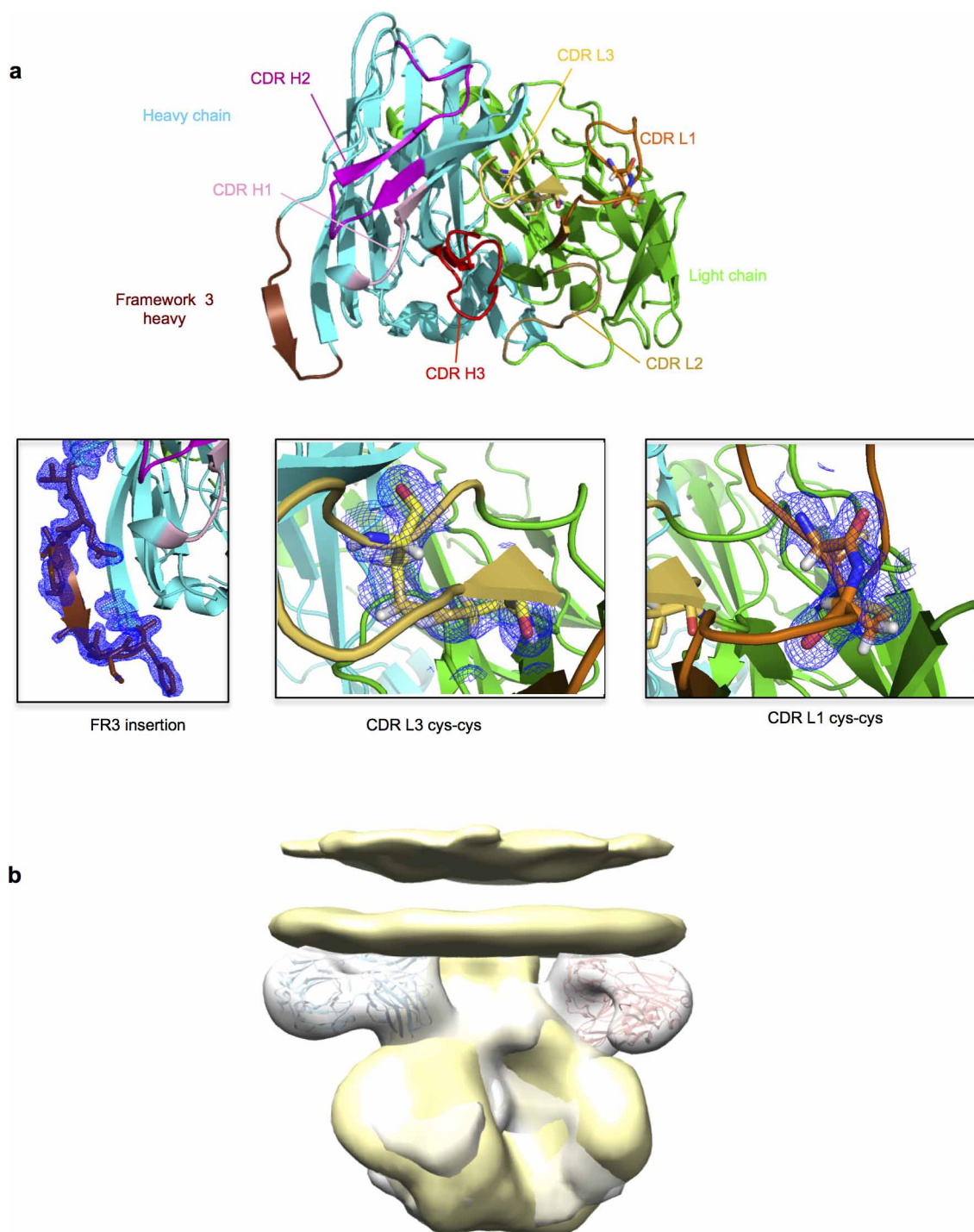
**Extended Data Figure 4 | Neutralization of 35O22 against pseudovirus mutants known to knock out activity against known glycan-specific antibodies.** **a**, Neutralization of 35O22 against JRCSF or KER2018.11 with or without the N160K mutation. PG9 and PG16 were used as positive controls.

**b**, Neutralization of 35O22 against N332A mutants of JRCSF. PGT121 was used as a positive control. **c**, Neutralization of 35O22 against N234S, T236K and N276D mutants of 3337.V2.C6. 8ANC195 was used as a positive control. Error bars denote one standard error of the mean (s.e.m.).



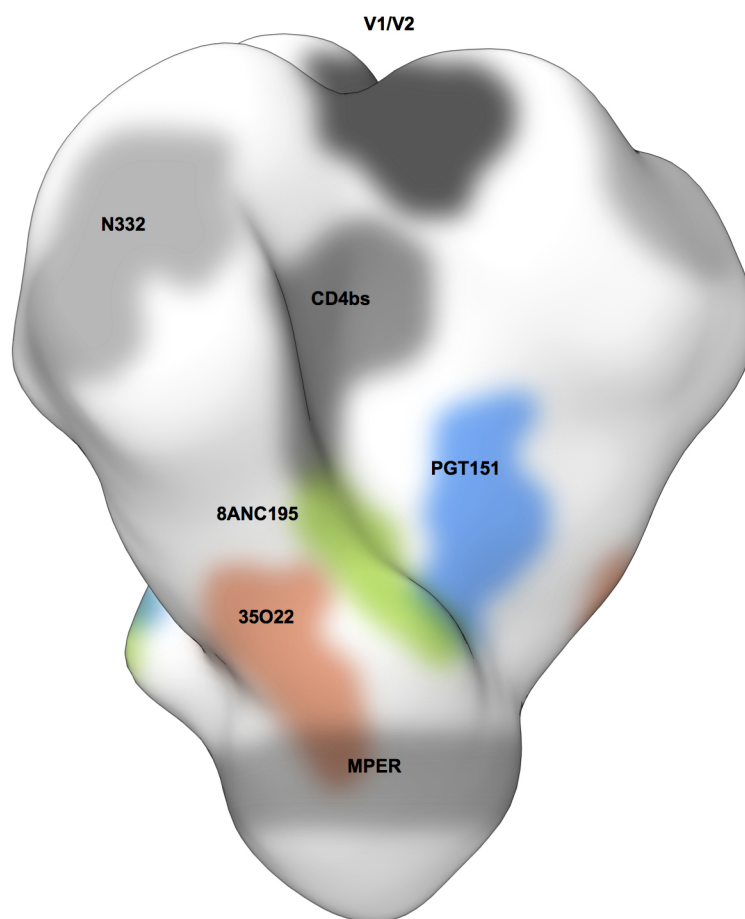


**Extended Data Figure 5 | Binding specificity of 35O22.** **a**, ELISA binding of indicated monoclonal antibodies to HIV<sub>YU2</sub> gp140 foldon trimer, gp120 and gp41 monomers. **b**, **c**, ELISA binding of gp120 (**b**) and gp140 (**c**) monomers from different HIV-1 subtypes.

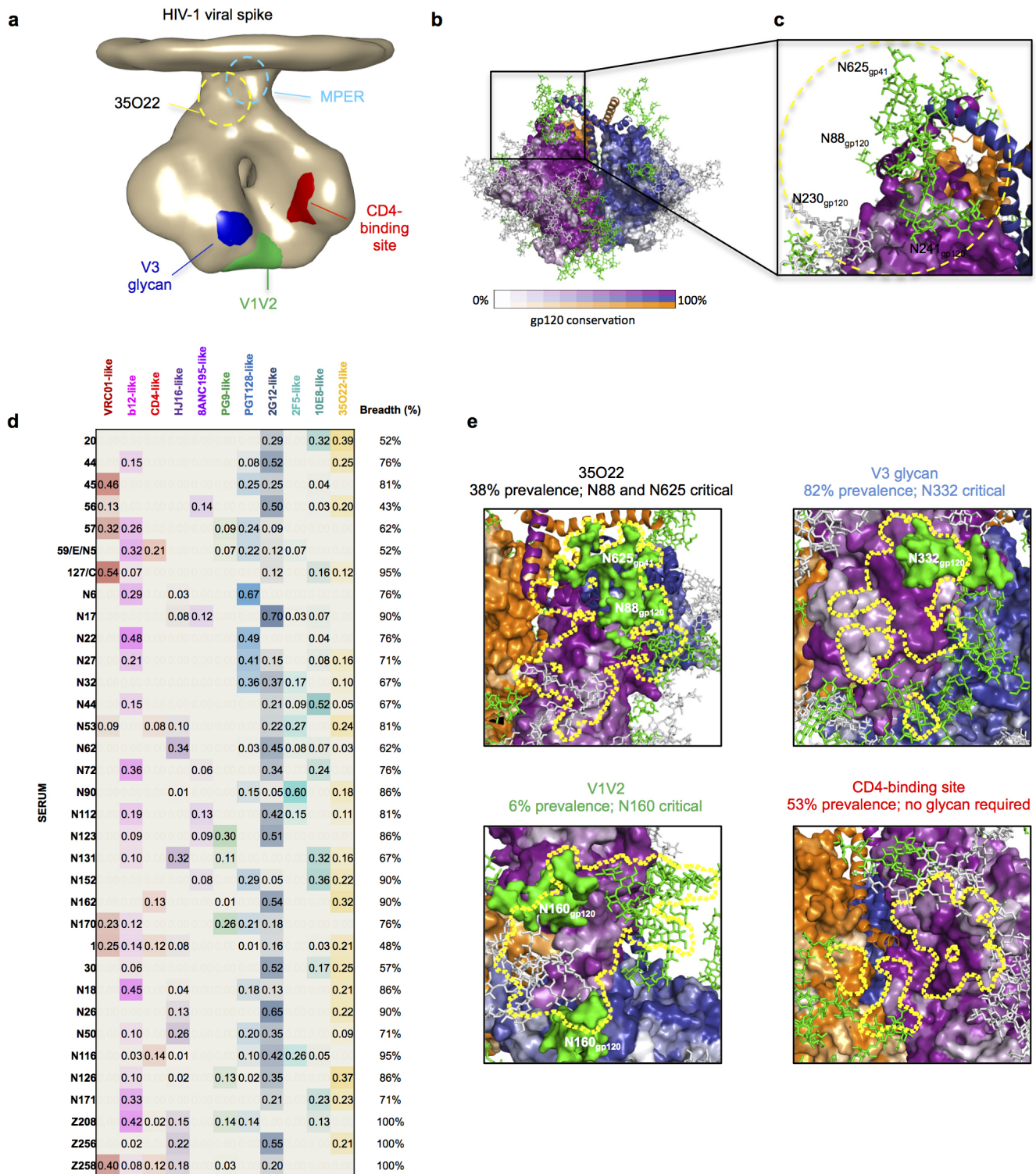


**Extended Data Figure 6 | 35O22 Fab features.** **a**, 35O22 is seen looking down on the combining site from the viewpoint of antigen in ribbons with the CDR coloured as in Fig. 3a. Insets (bottom row) show structural details of the framework 3 insertion and disulphides in CDR L1 and CDR L3 with electron

density  $2F_o - F_c$  contoured at  $1\sigma$ . **b**, Superposition of HIV<sub>BaL</sub> gp160 negative stain (yellow surface) with the negative stain reconstruction of soluble BG505 SOSIP in complex with 35O22 (grey surface) gives an estimation of the viral membrane location relative to 35O22 antibody as shown in Fig. 3b.



**Extended Data Figure 7 | Binding site of 35O22 on the HIV Env trimer.** Binding site of 35O22 (red) relative to those of PGT151 (blue) or 8ANC195 (green) are shown.



**Extended Data Figure 8 | A new site of HIV-1 vulnerability at the interface of gp120 and gp41 and prevalence of targeting.** **a**, Dominant sites of vulnerability to neutralizing antibody elicited by natural infection, shown in the context of an EM tomogram from the BaL viral spike. The viral membrane is positioned at the top of the spike. It is unclear if 35O22 and MPER antibodies bind to this form of the viral spike, and approximate locations for these are shown in dotted outlines. **b**, Viral spike from the soluble BG505 SOSIP context, shown in the same orientation as **a**, with gp120 surface coloured by conservation from 0% to 100%, from 4,265 HIV-1 strains (white to purple for protomer 1 with scale shown, white to blue for protomer 2, and white to orange for protomer 3), with glycans shown in green when present in more than 90% of strains, in grey when present in 30–90% of strains and not shown otherwise. **c**, 35O22-identified site of HIV-1 vulnerability comprises both conserved

amino acids and a cluster of glycans, including N88 from gp120 and N625 from gp41. N230 and N241 are not present in BG505 strain. The 35O22 epitope is shown by a yellow dotted line. **d**, Neutralization fingerprints for 35O22 and for antibodies encompassing ten different epitope specificities representing the other four known major sites of Env vulnerability were used to interrogate the serum specificities of 34 HIV-infected patients. Values (with proportional colour intensities) predict the fraction of serum neutralization that can be attributed to each antibody specificity. Possible 35O22-like signals were predicted for 13 of the sera (values >0.2), while strong signals were observed in 3 of the sera (values >0.3). A panel of 21 HIV-1 strains was used in the neutralization analysis and for computing serum breadth. **e**, Sites of HIV-1 vulnerability to neutralizing antibody outlined by a yellow line. Prevalence in a 34-donor cohort and critical glycans are indicated.



[illegible]

**Extended Data Figure 9 | Autologous virus Env sequences and the impact of variants on 35O22 neutralization.** a, A total of 12 single-genome amplicons from plasma of patient N152 were sequenced. Donor Env sequences together with the reference sequences of JRCSF and LAI are aligned. Amino acids critical for 35O22 neutralization of JRCSF and LAI are labelled in yellow.

Differences between autologous and JRCSF sequences are labelled in green.  
b, 35O22 neutralization of JRCSF pseudovirus or variants containing the autologous virus mutations from patient N152. Error bars denote one standard error of the mean.

Extended Data Table 1 | Neutralizing activity of sera or monoclonal antibodies against HIV<sub>JRCSF</sub> pseudovirus with mutation in the 35022 epitope

Sera ID	ID50					Fold change*			
	WT	N88A	N230A	N241A	N625A	N88A	N230A	N241A	N625A
20	6864	7396	5295	5261	5781	0.9	1.3	1.3	1.2
44	6095	2824	1177	2632	776	2.2	5.2	2.3	7.9
45	4019	2422	2094	1678	3265	1.7	1.9	2.4	1.2
56	655	1124	2017	89	1901	0.6	0.3	7.4	0.3
57	7586	626	4981	393	355	12.1	1.5	19.3	21.4
127/C	5217	820	1829	1082	789	6.4	2.9	4.8	6.6
N6	5966	214	437	436	679	27.9	13.7	13.7	8.8
N17	62354	1121	743	2279	1653	55.6	83.9	27.4	37.7
N22	9735	932	3017	1591	3523	10.4	3.2	6.1	2.8
N32	3131	539	1238	2538	686	5.8	2.5	1.2	4.6
N44	5439	1532	2110	2622	2003	3.6	2.6	2.1	2.7
N53	6690	963	1048	352	1044	6.9	6.4	19.0	6.4
N72	6722	1484	1058	199	2482	4.5	6.4	33.8	2.7
N90	4275	3853	1479	1018	10576	1.1	2.9	4.2	0.4
N112	5903	2344	967	706	6071	2.5	6.1	8.4	1.0
N123	21392	1194	1217	771	824	17.9	17.6	27.7	26.0
N131	18777	106	402	163	275	177.1	46.7	115.2	68.3
N152	36636	6677	28758	4339	4279	5.5	1.3	8.4	8.6
N162	17492	558	4078	718	1739	31.3	4.3	24.4	10.1
N170	16472	1342	6564	990	448	12.3	2.5	16.6	36.8
1	3485	1605	343	197	256	2.2	10.2	17.7	13.6
30	14255	5620	3314	13125	6221	2.5	4.3	1.1	2.3
N26	19628	1853	2443	1951	3917	10.6	8.0	10.1	5.0
N18	60183	580	1202	480	656	103.8	50.1	125.4	91.7
N116	586	2504	2176	1948	616	0.2	0.3	0.3	1.0
N27	1088	1446	4656	838	790	0.8	0.2	1.3	1.4
N62	907	536	462	429	263	1.7	2.0	2.1	3.4
N126	2502	1132	1146	3554	1807	2.2	2.2	0.7	1.4
N171	12329	5967	3125	1948	1904	2.1	3.9	6.3	6.5
Z208	21501	9337	4338	8740	1815	2.3	5.0	2.5	11.8
Z256	11467	1584	1881	2055	1240	7.2	6.1	5.6	9.2
Z258	31647	3154	2018	2094	4114	10.0	15.7	15.1	7.7
mAb	IC50					Fold change†			
10E8	0.078	0.0098	0.01984	0.04513	0.01757	0.13	0.3	0.58	0.23
PGT121	0.01508	0.02161	0.02053	0.02399	0.02431	1.43	1.4	1.59	1.61
PG9	0.00692	0.00216	0.0022	0.00189	0.00238	0.31	0.3	0.27	0.34
35O22	0.0106	>20	>20	>20	>20	>1886	>1886	>1886	>1886

\*Fold change indicates ID<sub>50</sub> of HIV<sub>JRCSF</sub> WT/ID<sub>50</sub> of HIV<sub>JRCSF</sub> mutant. Values with fold changes >5 are highlighted in yellow.†Fold change indicates IC<sub>50</sub> of HIV<sub>JRCSF</sub> mutant/IC<sub>50</sub> of HIV<sub>JRCSF</sub> WT. Values with fold changes >5 are highlighted in yellow.

# Pseudouridine profiling reveals regulated mRNA pseudouridylation in yeast and human cells

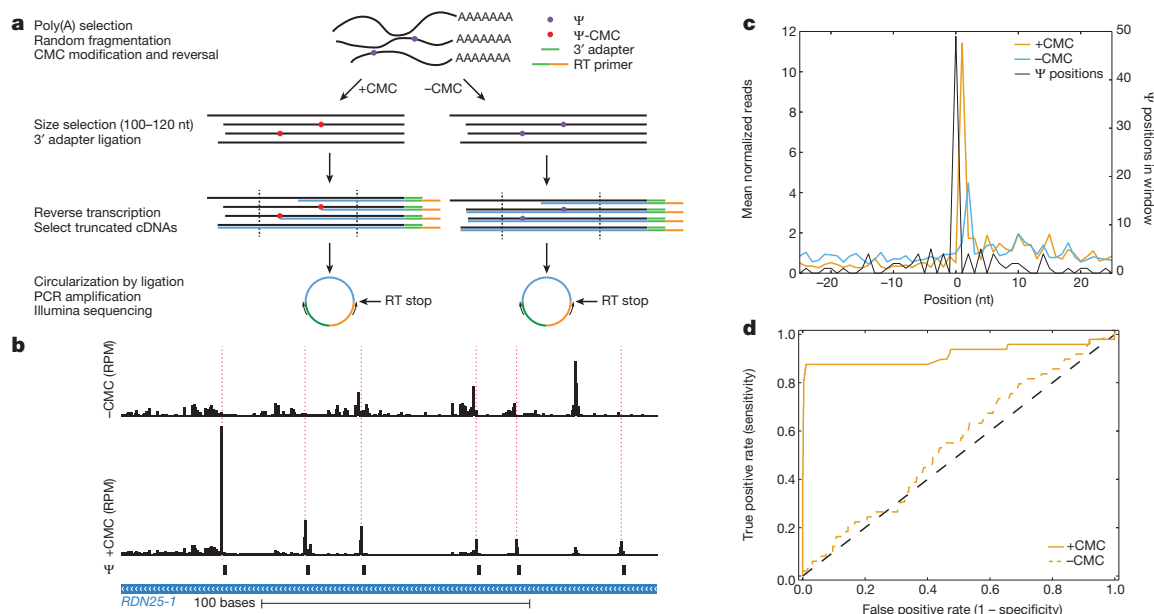
Thomas M. Carlile<sup>1</sup>, Maria F. Rojas-Duran<sup>1</sup>, Boris Zinshteyn<sup>1</sup>, Hakyung Shin<sup>1</sup>, Kristen M. Bartoli<sup>1</sup> & Wendy V. Gilbert<sup>1</sup>

Post-transcriptional modification of RNA nucleosides occurs in all living organisms. Pseudouridine, the most abundant modified nucleoside in non-coding RNAs<sup>1</sup>, enhances the function of transfer RNA and ribosomal RNA by stabilizing the RNA structure<sup>2–8</sup>. Messenger RNAs were not known to contain pseudouridine, but artificial pseudouridylation dramatically affects mRNA function—it changes the genetic code by facilitating non-canonical base pairing in the ribosome decoding centre<sup>9,10</sup>. However, without evidence of naturally occurring mRNA pseudouridylation, its physiological relevance was unclear. Here we present a comprehensive analysis of pseudouridylation in *Saccharomyces cerevisiae* and human RNAs using Pseudo-seq, a genome-wide, single-nucleotide-resolution method for pseudouridine identification. Pseudo-seq accurately identifies known modification sites as well as many novel sites in non-coding RNAs, and reveals hundreds of pseudouridylated sites in mRNAs. Genetic analysis allowed us to assign most of the new modification sites to one of seven conserved pseudouridine synthases, Pus1–4, 6, 7 and 9. Notably, the majority of pseudouridines in mRNA are regulated in response to environmental signals, such as nutrient deprivation in yeast and serum starvation in human cells. These results suggest a mechanism for the rapid and regulated rewiring of the genetic code through inducible mRNA

modifications. Our findings reveal unanticipated roles for pseudouridylation and provide a resource for identifying the targets of pseudouridine synthases implicated in human disease<sup>11–13</sup>.

Although more than 100 classes of RNA modifications have been characterized, primarily in tRNA and rRNA<sup>14</sup>, only three modified nucleotides have been identified within the coding sequences of mRNA, N<sup>6</sup>-methyladenosine (m<sup>6</sup>A), 5-methylcytosine (m<sup>5</sup>C) and inosine<sup>15–19</sup>. To define the global landscape of RNA pseudouridylation *in vivo* and determine whether mRNAs contain pseudouridine (Ψ), we developed a high-throughput method to identify Ψ in the transcriptome with single-nucleotide resolution. Ψ can be selectively modified with *N*-cyclohexyl-*N'*-(2-morpholinoethyl)-carbodiimide metho-p-toluenesulphonate (CMC) to generate a block to reverse transcriptase one nucleotide 3' to the pseudouridylated site<sup>20</sup>. We exploited this chemistry to determine the locations of Ψ using next-generation sequencing (Fig. 1a; see Methods). Mock-treated (–CMC) RNA fragments were processed in parallel to identify pseudouridine-independent reverse transcription stops.

Using Pseudo-seq and stringent Ψ-calling criteria, we identified 42/51 known Ψs in rRNA and small nuclear RNA (Supplementary Table 1) with an observed false positive rate of 0.1%. The estimated false discovery rate (FDR) ranges from approximately 5% for highly expressed genes



**Figure 1 | Genome-wide pseudouridine sequencing with single-nucleotide resolution.** **a**, Schematic of Pseudo-seq library preparation. nt, nucleotides. **b**, Genome browser view of Pseudo-seq reads mapping to a 200-nt region of *RDN25-1* (chrXII: 452168–452367) containing six known Ψs, generated from pooled reads for  $n = 12$  technical replicates from wild-type log-phase yeast cultures. Peaks of Ψ-dependent reads are indicated with dashed red lines. Reads

per million (RPM). **c**, A metaPsi plot of mean normalized reads (left axis) for +CMC (orange) and –CMC (blue) libraries. The number of Ψs at each position in the metaPsi window is indicated (black, right axis). CMC-dependent reverse transcription stops are found 1 nt 3' of known Ψs. **d**, A receiver operating characteristic curve of the Pseudo-seq signal for all known Ψs in rRNA and U2 snRNA.

<sup>1</sup>Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA.



to approximately 12.5% for lowly expressed genes. (Fig. 1b–d and Methods). Six out of nine false negatives were due to ‘shadowing’ from reverse transcription stops 3’ of the  $\Psi$  (for example, 25S- $\Psi$ 2258 was not detected upstream of  $\Psi$ 2260). We also identified many  $\Psi$ s in tRNA, all of which occurred at known positions (Supplementary Table 2). We verified the single-nucleotide resolution of Pseudo-seq by profiling four small nucleolar RNA deletion mutants that eliminate pseudouridylation of nine specific rRNA and snRNA target sites (Extended Data Fig. 1a–d). Similar specificity and sensitivity were achieved using different reverse transcription enzymes, RNA fragment lengths, CMC concentrations, and truncated cDNA lengths, demonstrating the robustness of the Pseudo-seq method (Extended Data Fig. 2a–d).

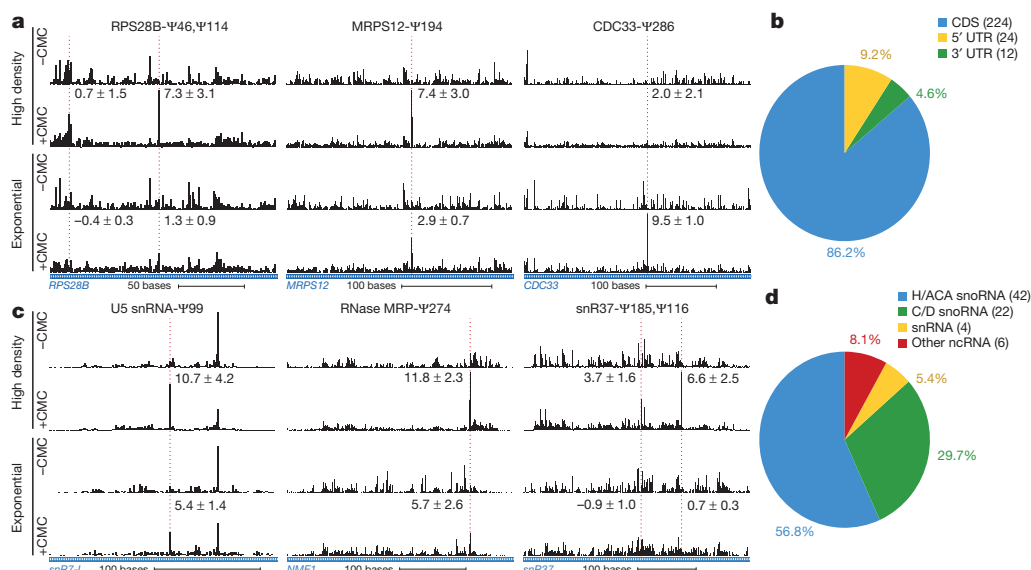
After validating the ability of Pseudo-seq to detect known  $\Psi$ s in non-coding RNAs, we next analysed mRNA pseudouridylation in budding yeast during post-diauxic growth ( $A_{600\text{ nm}} = 12$ ) (Extended Data Fig. 3a). To define high confidence Pseudo-seq hits even in transcripts with sparse read coverage, we required reproducibility in 10/14 independent experiments (Extended Data Fig. 3b, c). Notably, we found that many mRNAs contain  $\Psi$  (Fig. 2a). In total, we conservatively identified 260  $\Psi$ s in 238 protein-coding transcripts (Supplementary Table 3). Relaxing our criteria to include  $\Psi$ s detected in 9/14 experiments, a category that includes the known  $\Psi$ 56 in U2 snRNA, increased the number of candidate mRNA  $\Psi$ s to 466. We established a rough detectability threshold by determining the lowest observed expression level of genes having sufficient reads for reproducible  $\Psi$  calling; 5,278 genes passed the cutoff. Thus, it is unlikely that there are substantially more mRNA  $\Psi$ s to be discovered in post-diauxic yeast. We conclude that mRNA pseudouridines are relatively scarce.  $\Psi$ s were found in 5’ transcript leaders (5’TLs), coding sequences (CDS), and 3’ untranslated regions (3’ UTRs) with an underrepresentation of  $\Psi$  in 3’ UTRs ( $P = 10^{-4}$ , hypergeometric test) (Fig. 2b). GUA valine codons were the most frequently modified, suggesting the existence of a sequence-specific mechanism for mRNA pseudouridylation (Extended Data Fig. 4).

We investigated the potential for regulation of mRNA pseudouridylation by comparing two cellular conditions with substantial differences in gene expression and physiology: log phase and post-diauxic growth. Remarkably, most mRNA  $\Psi$ s were regulated: 42% of the sites modified during post-diauxic growth were not detectably modified in log phase,

whereas other sites, such as CDC33  $\Psi$ 286, were much more extensively modified during exponential growth. Moreover, of the 150 modified sites detected in both log phase and post-diauxic growth, 62 showed >twofold changes in peak height between conditions indicating growth-state-dependent changes in the extent of mRNA modification (Fig. 2a and Supplementary Table 3). Importantly, we ruled out differences in mRNA expression as an explanation for condition-dependent differences in  $\Psi$  detection (Extended Data Fig. 5). Thus, the process of mRNA pseudouridylation is regulated in response to environmental cues.

Yeast non-coding RNAs (ncRNA) have been extensively characterized for post-transcriptional modifications. Nevertheless, we identified 74 novel pseudouridylated sites in ncRNAs (Supplementary Table 4). A few, like  $\Psi$ 274 in the RNase MRP RNA (*NME1*) (Fig. 2c), were constitutively modified, while most, including the previously described  $\Psi$ 56 and  $\Psi$ 93 in U2 snRNA<sup>21</sup>, were induced during post-diauxic growth (Extended Data Fig. 6a). snoRNAs were notably enriched among ncRNA classes with regulated pseudouridines: 19/29 H/ACA and 14/47 C/D snoRNAs showed one or more sites specifically modified in cells grown to high density (Fig. 2d, Extended Data Fig. 6b, c). Pseudouridylation of rRNA sites changed very little: only one site, 25S- $\Psi$ 2314, changed more than twofold. However, owing to the stability of rRNA and the greatly reduced rate of ribosome synthesis during post-diauxic growth, we cannot rule out production of a minority population of differentially modified ribosomes in dense cultures.

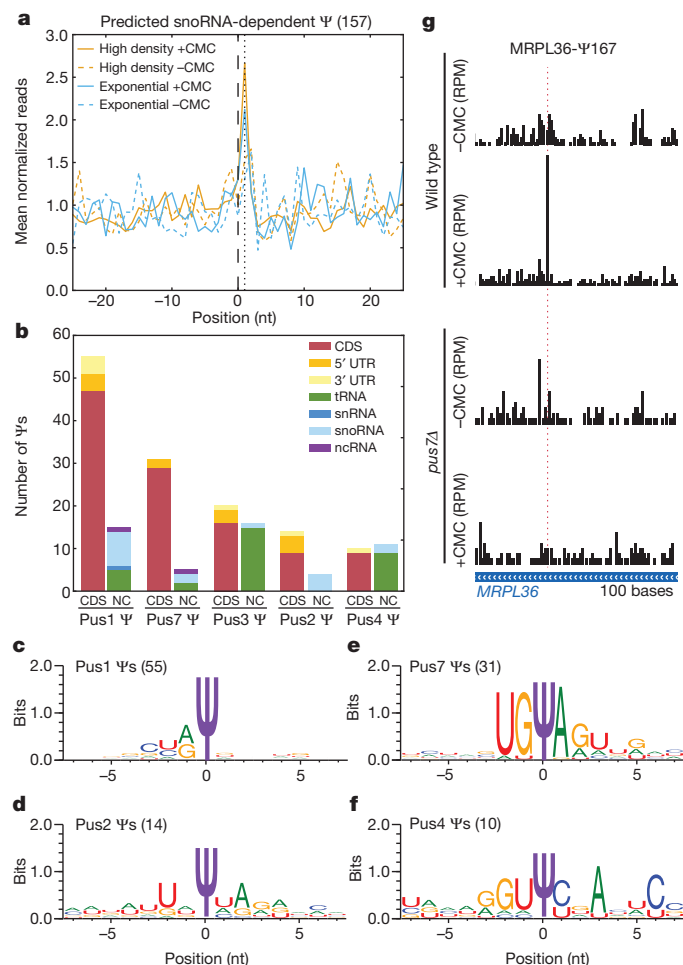
We next sought to define the molecular basis for targeting of novel mRNA and ncRNA sites for pseudouridylation.  $\Psi$ s in rRNA, snRNA and tRNA are produced by two classes of enzymes with distinct modes of target recognition. The first class, which includes yeast Cbf5 and human dyskerin, associates with H/ACA snoRNAs to direct pseudouridylation of sites that base pair with the snoRNA guide sequences, while the second class recognizes its targets without the aid of an RNA guide. We computationally identified 157 unique sites in mRNAs containing perfect matches to canonical snoRNA targets (Supplementary Table 5). When these potential pseudouridylation sites were considered in aggregate, statistically significant pseudouridylation was detected (Fig. 3a, Extended Data Fig. 7a, b), which increased with the number of base pairs to the snoRNA guide sequence and was specific to post-diauxic growth (Extended Data Fig. 7c, d). However, only three such sites passed our threshold



**Figure 2 | Yeast mRNAs and ncRNAs are inducibly pseudouridylated.** **a**, **c**, CMC-dependent peaks of reads are indicated with a dashed red line. The median Pseudo-seq peak heights in each condition are given  $\pm$  s.d.; negative peak values occur when the reads in the –CMC library exceed those in the +CMC library. Traces are representative of four wild-type biological replicates. **a**, Pseudo-seq reads in *RPS28B* (chrXII: 673163–673336), *MRPS12* (chrXIV:

694489–694736) and *CDC33* (chrXV: 50560–60875). **b**, Summary of locations of  $\Psi$ s within mRNA features. CDS, coding sequence. **c**, Pseudo-seq reads in U5 snRNA (*snR7-L*, chrVII: 939458–939671), RNase MRP RNA (*NME1*, chrXIV:585585–585925) and an H/ACA snoRNA (*snR37*, chrX: 228090–228475). **d**, Summary of novel  $\Psi$ s identified in ncRNA.





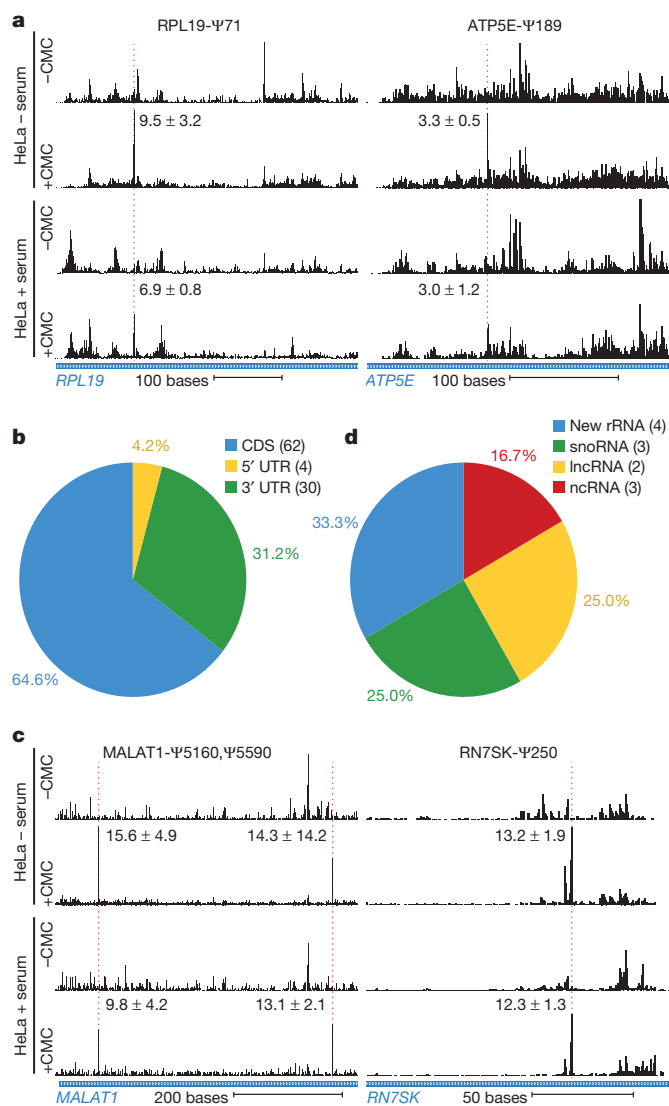
**Figure 3 | Mechanisms of mRNA pseudouridylation.** **a**, MetaPsi plot of mean normalized reads for computationally predicted snoRNA-dependent targets in mRNA from high density cultures (orange), log phase cultures (blue), +CMC (solid) and -CMC (dashed). Indicated are the predicted snoRNA target site (black, dashed), and the expected peak of CMC-dependent reads (black, dotted). Reads were pooled from four wild-type biological replicate libraries. **b**, Summary of  $\Psi$ s identified by Pseudo-seq as *PUS*-dependent. The few *PUS6*- and *PUS9*-dependent  $\Psi$ s are not shown. The locations of  $\Psi$ s within mRNAs and the distribution of  $\Psi$ s among ncRNA types are indicated. NC, noncoding sequence. **c-f**, Sequence motifs surrounding *PUS1*- (**c**), *PUS2*- (**d**), *PUS7*- (**e**) and *PUS4*-dependent (**f**)  $\Psi$ s in mRNAs, generated by WebLogo 3.3. **g**, *PUS7*-dependent Pseudo-seq reads in *MRPL36* (chrII: 484301-484400).

for  $\Psi$  calling on their own (Extended Data Fig. 7e). Thus, it is probable that many additional mRNAs are pseudouridylated at a low level and our estimate of 260 mRNA  $\Psi$ s represents a conservative minimum.

Because most pseudouridylated sites showed no significant complementarity to snoRNA guide sequences, we next investigated whether snoRNA-independent pseudouridine synthases are responsible for modifying sites in mRNAs. Yeast has nine pseudouridine synthase (*PUS*) genes, all of which are expressed in both log phase and post-diauxic growth. We profiled the eight viable *PUS* deletion strains (*pusΔ*) grown to high density and identified mRNA targets for each *Pus* protein, with the exception of *Pus5* whose only known target is the 21S mitochondrial rRNA<sup>22</sup> (Fig. 3b, Extended Data Fig. 8a, b and Supplementary Table 6). The largest number of mRNA and novel ncRNA Ψs could be assigned to *Pus1*, a member of the TruA family that constitutively modifies multiple positions in cytoplasmic tRNAs and one position in U2 snRNA by a mode of target recognition that is incompletely defined. Whereas known *Pus1*-dependent tRNA targets showed constitutive pseudouridylation as expected, most of the mRNA targets showed increased

modification during post-diauxic growth (Extended Data Fig. 8c, Supplementary Table 3). The mRNA targets of Pus1 showed little similarity at the primary sequence level, consistent with the proposed structure-dependent mode of target recognition by this enzyme (Fig. 3c, Extended Data Fig. 8d)<sup>23</sup>, while Pus2, a close paralogue of Pus1, had 14 mRNA targets with a weak sequence consensus distinct from Pus1 (Fig. 3d, Extended Data Fig. 8e). Intriguingly, the Pus1 targets included seven genes encoding five proteins of the large ribosomal subunit, a significant enrichment ( $P = 0.025$ ). Our comprehensive pseudouridine profiling more than doubles the number of known substrates of Pus1 and Pus2, identifies unanticipated mRNA targets, and provides the first demonstration of regulated pseudouridylation by these enzymes.

Unlike Pus1 and Pus2, the mRNA targets of Pus4 and Pus7 contained clear consensus sites in agreement with the known sequence requirements for these enzymes to modify their canonical tRNA targets, UGΨAR for Pus7 and GUΨCNANNC for Pus4 (Fig. 3e–g, Extended Data



**Figure 4 | Regulated pseudouridylation of human RNAs.** Pseudo-seq was performed on HeLa cells grown in the presence or absence of serum for 24 h. **a, b**, CMC-dependent peaks of reads are indicated with a dashed red line. The median Pseudo-seq peak heights in each condition are given  $\pm$  s.d. Traces are representative of  $n = 4$  (–serum), and  $n = 5$  (+serum) biological replicates. Genome browser views represent spliced transcripts. **a**, Pseudo-seq reads from *RPL19* (12–460), and *ATP5E* (154–437). **b**, Summary of locations of  $\Psi$ s within mRNA features. **c**, Pseudo-seq reads from *MALAT1* (5081–5636) and *RN7SK* (142–307). **d**, Summary of novel  $\Psi$ s identified in ncRNA.

Fig. 8f–h)<sup>24,25</sup>. We also identified novel targets for Pus3 (20 mRNA, 1 ncRNA), Pus6 (3, 1) and Pus9 (1, 0), and, in total, assigned 52% of mRNA  $\Psi$ s and 31% of novel ncRNA  $\Psi$ s to individual Pus proteins. The remaining sites may be modified by the essential protein Pus8 and/or may be redundantly targeted by multiple Pus proteins. Together, these results reveal unanticipated diversity in Pus targets and show that Pus-dependent non-tRNA sites are regulated in response to changing cellular growth conditions. The discovery of novel mRNA substrates for Pus proteins raises the possibility that other tRNA modifying enzymes may likewise target mRNAs.

As the pseudouridine synthases that modify yeast mRNAs are conserved throughout eukaryotes, we investigated whether regulated mRNA pseudouridylation also occurs in mammalian cells. Human cervical carcinoma (HeLa) cells were profiled during normal proliferation and 24 h after serum starvation. Pseudo-seq detected known pseudouridines with good sensitivity and specificity (Supplementary Table 7, Extended Data Fig. 9a–c). By restricting our analysis to more highly expressed genes and requiring reproducibility in four independent biological replicates, we conservatively identified 96  $\Psi$ s in 89 human mRNAs (Supplementary Table 8). As in yeast, some  $\Psi$  modifications in human mRNAs were regulated by cellular growth state (Fig. 4a, Extended Data Fig. 10a, b), and modified sites were found throughout the transcript (Fig. 4b). We also discovered novel  $\Psi$ s in human ncRNAs, including 4 previously unknown sites in rRNA (Extended Data Fig. 10c, Supplementary Table 9) and sites in long non-coding (lnc-), sn- and snoRNAs (Fig. 4c, d). Thus, the Pseudo-seq approach is broadly applicable to diverse organisms and growth states. Moreover, the phenomenon of regulated mRNA pseudouridylation is conserved from yeast to humans.

In summary, Pseudo-seq provides comprehensive analysis of RNA pseudouridylation with single-nucleotide resolution and reveals that endogenous mRNAs are specifically pseudouridylated in a highly regulated manner in yeast and human cells. Because  $\Psi$  stabilizes RNA structure, mRNA pseudouridylation could alter translation initiation efficiency<sup>26,27</sup>, ribosome pausing<sup>28</sup>, RNA localization<sup>29</sup> and regulation by RNA interference<sup>30</sup>, to name a few aspects of mRNA metabolism known to be affected by RNA structure, although we cannot exclude the possibility that many instances of mRNA pseudouridylation may be functionally silent. However, given recent evidence that pseudouridine profoundly affects decoding by ribosomes from diverse organisms<sup>3</sup>, our results also raise the possibility of widespread regulated rewiring of the genetic code. Finally, this work suggests that diseases associated with mutations in pseudouridine synthases, including mitochondrial myopathy and sideroblastic anaemia (MLASA)<sup>11</sup>, dyskeratosis congenita<sup>12</sup> and lung cancer<sup>13</sup>, could be due to misregulation of mRNA targets.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 27 May; accepted 27 August 2014.**

**Published online 5 September 2014.**

1. Davis, F. F. & Allen, F. W. Ribonucleic acids from yeast which contain a fifth nucleotide. *J. Biol. Chem.* **227**, 907–915 (1957).
2. Arnez, J. G. & Steitz, T. A. Crystal structure of unmodified tRNA<sup>Gln</sup> complexed with glutamyl-tRNA synthetase and ATP suggests a possible role for pseudo-uridines in stabilization of RNA structure. *Biochemistry* **33**, 7560–7567 (1994).
3. Charette, M. & Gray, M. W. Pseudouridine in RNA: what, where, how, and why. *IUBMB Life* **49**, 341–351 (2000).
4. Davis, D. R. & Poulter, C. D. <sup>1</sup>H–<sup>15</sup>N NMR studies of *Escherichia coli* tRNA<sup>Phe</sup> from *hisT* mutants: a structural role for pseudouridine. *Biochemistry* **30**, 4223–4231 (1991).
5. Davis, D. R., Veltri, C. A. & Nielsen, L. An RNA model system for investigation of pseudouridine stabilization of the codon-anticodon interaction in tRNA<sup>Lys</sup>, tRNA<sup>His</sup> and tRNA<sup>Tyr</sup>. *J. Biomol. Struct. Dyn.* **15**, 1121–1132 (1998).
6. Hall, K. B. & McLaughlin, L. W. Properties of a U1/mRNA 5' splice site duplex containing pseudouridine as measured by thermodynamic and NMR methods. *Biochemistry* **30**, 1795–1801 (1991).

7. Hudson, G. A., Bloomingdale, R. & Znosko, B. Thermodynamic contribution and nearest-neighbor parameters of pseudouridine-adenosine base pairs in oligoribonucleotides. *RNA* **19**, 1474–1482 (2013).
8. Yarian, C. S. *et al.* Structural and functional roles of the N1- and N3-protons of psi at tRNA's position 39. *Nucleic Acids Res.* **27**, 3543–3549 (1999).
9. Fernández, I. S. *et al.* Unusual base pairing during the decoding of a stop codon by the ribosome. *Nature* **500**, 107–110 (2013).
10. Karjilovich, J. & Yu, Y.-T. Converting nonsense codons into sense codons by targeted pseudouridylation. *Nature* **474**, 395–398 (2011).
11. Bykhovskaya, Y., Casas, K., Mengesha, E., Inbal, A. & Fischel-Ghodsian, N. Missense mutation in pseudouridine synthase 1 (PUS1) causes mitochondrial myopathy and sideroblastic anemia (MLASA). *Am. J. Hum. Genet.* **74**, 1303–1308 (2004).
12. Heiss, N. S. *et al.* X-linked dyskeratosis congenita is caused by mutations in a highly conserved gene with putative nucleolar functions. *Nature Genet.* **19**, 32–38 (1998).
13. Mei, Y.-P. *et al.* Small nucleolar RNA 42 acts as an oncogene in lung tumorigenesis. *Oncogene* **31**, 2794–2804 (2012).
14. Cantara, W. A. *et al.* The RNA Modification Database, RNAMDB: 2011 update. *Nucleic Acids Res.* **39**, D195–D201 (2011).
15. Chen, L. Characterization and comparison of human nuclear and cytosolic editases. *Proc. Natl Acad. Sci. USA* **110**, E2741–E2747 (2013).
16. Dominissini, D. *et al.* Topology of the human and mouse m<sup>6</sup>A RNA methylomes revealed by m<sup>6</sup>A-seq. *Nature* **485**, 201–206 (2012).
17. Li, J. B. *et al.* Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. *Science* **324**, 1210–1213 (2009).
18. Meyer, K. D. *et al.* Comprehensive analysis of mRNA methylation reveals enrichment in 3' UTRs and near stop codons. *Cell* **149**, 1635–1646 (2012).
19. Squires, J. E. *et al.* Widespread occurrence of 5-methylcytosine in human coding and non-coding RNA. *Nucleic Acids Res.* **40**, 5023–5033 (2012).
20. Bakin, A. & Ofengand, J. Four newly located pseudouridylate residues in *Escherichia coli* 23S ribosomal RNA are all at the peptidyltransferase center: analysis by the application of a new sequencing technique. *Biochemistry* **32**, 9754–9762 (1993).
21. Wu, G., Xiao, M., Yang, C. & Yu, Y.-T. U2 snRNA is inducibly pseudouridylated at novel sites by Pus7p and snR81 RNP. *EMBO J.* **30**, 79–89 (2011).
22. Ansmant, I., Massenot, S., Grosjean, H., Motorin, Y. & Branlant, C. Identification of the *Saccharomyces cerevisiae* RNA:pseudouridine synthase responsible for formation of  $\Psi$ <sub>2819</sub> in 21S mitochondrial ribosomal RNA. *Nucleic Acids Res.* **28**, 1941–1946 (2000).
23. Arluison, V., Buckle, M. & Grosjean, H. Pseudouridine synthetase Pus1 of *Saccharomyces cerevisiae*: kinetic characterisation, tRNA structural requirement and real-time analysis of its complex with tRNA. *J. Mol. Biol.* **289**, 491–502 (1999).
24. Becker, H. F., Motorin, Y., Sissler, M., Florentz, C. & Grosjean, H. Major identity determinants for enzymatic formation of ribothymidine and pseudouridine in the T $\Psi$ -loop of yeast tRNAs. *J. Mol. Biol.* **274**, 505–518 (1997).
25. Behm-Ansmant, I. *et al.* The *Saccharomyces cerevisiae* U2 snRNA:pseudouridine-synthase Pus7p is a novel multisite-multisubstrate RNA: $\Psi$ -synthase also acting on tRNAs. *RNA* **9**, 1371–1382 (2003).
26. Kudla, G., Murray, A. W., Tollervey, D. & Plotkin, J. B. Coding-sequence determinants of gene expression in *Escherichia coli*. *Science* **324**, 255–258 (2009).
27. Shah, P., Ding, Y., Niemczyk, M., Kudla, G. & Plotkin, J. B. Rate-limiting steps in yeast protein translation. *Cell* **153**, 1589–1601 (2013).
28. Somogyi, P., Jenner, A. J., Brierley, I. & Inglis, S. C. Ribosomal pausing during translation of an RNA pseudoknot. *Mol. Cell. Biol.* **13**, 6931–6940 (1993).
29. Jambhekar, A. & Derisi, J. L. Cis-acting determinants of asymmetric, cytoplasmic RNA transport. *RNA* **13**, 625–642 (2007).
30. Tan, X. *et al.* Tiling genomes of pathogenic viruses identifies potent antiviral shRNAs and reveals a role for secondary structure in shRNA efficacy. *Proc. Natl Acad. Sci. USA* **109**, 869–874 (2012).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** We thank I. Cheeseman, C. Burge, U. RajBhandary, D. Bartel and members of the Gilbert laboratory for comments and discussion. The sequencing was performed at the BioMicro Center under the direction of S. Levine. This work was supported by grants from The American Cancer Society – Robbie Sue Mudd Kidney Cancer Research Scholar Grant (RSG-13-396-01-RMC) and the National Institutes of Health (GM094303, GM081399) to W.V.G. T.M.C. was supported by the American Cancer Society New England Division (Ellison Foundation Postdoctoral Fellowship), and K.M.B. was supported by a Postdoctoral Fellowship (PF-13-319-01 – RMC) from the American Cancer Society. This work was supported in part by the NIH Pre-Doctoral Training Grant T32GM007287.

**Author Contributions** T.M.C. and W.V.G. conceived and designed the experiments. T.M.C., M.F.R.-D., H.S., K.M.B. and W.V.G. performed the experiments. T.M.C., B.Z. and H.S. performed the bioinformatic analyses. T.M.C. and W.V.G. interpreted the results and wrote the paper with input from all authors.

**Author Information** Data have been deposited in NCBI's Gene Expression Omnibus (GEO), accession number GSE58200. Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to W.V.G. ([wgilbert@mit.edu](mailto:wgilbert@mit.edu)).

## METHODS

**Yeast strains and growth.** All yeast strains are *Saccharomyces cerevisiae* BY4741 or BY4742 derivatives (BY4742: wild type (YWG11), *snr37A* (YWG287, YWG343), *snr43A* (YWG293), *snr49A* (YWG299, YWG354), *snr81A* (YWG322, YWG376), *pus4A* (YWG1251, YWG1252), BY4741: *pus1A* (YWG1209), *pus1A* (YWG1209), *pus2A* (YWG1210), *pus3A* (YWG1211), *pus5A* (YWG1212), *pus6A* (YWG1213), *pus7A* (YWG1214), *pus9A* (YWG1215)). The snoRNA deletion strains and *pus4A* strains were made using PCR-based deletion cassettes<sup>31</sup>. The other *pusA* strains were obtained from the Yeast Deletion Collection<sup>32</sup>. Strains were grown at 30 °C in YPAD (1% yeast extract, 2% peptone, 0.01% adenine hemisulphate, 2% glucose), and were harvested by centrifugation in log phase ( $A_{600\text{ nm}} \approx 1$ ), or at high density ( $A_{600\text{ nm}} \approx 12-15$ ).

**Cell culture.** HeLa (human cervix adenocarcinoma; CCL-2, ATCC) cells were cultured in DMEM (D6429; Sigma) supplemented with 10% fetal bovine serum (FBS; Atlanta Biologicals). Cells were grown at 37 °C with 5% CO<sub>2</sub> under standard laboratory conditions. For serum starvation cells were plated at a density of  $5 \times 10^6$  per 150-mm plate in DMEM+10% FBS, 24 h before the experiment. Cells were then washed three times in PBS, before the addition of either serum-free medium (DMEM, no FBS) or full medium containing FBS (DMEM+10% FBS) for 24 h.

**Pseudo-seq library preparation.** Yeast total RNA was isolated by hot acid phenol extraction, followed by isopropanol precipitation<sup>33</sup>. HeLa total RNA was isolated using QIAzol (Qiagen; 79306). PolyA+ RNA was isolated from 10 mg (yeast) or 2 mg (HeLa) total RNA using oligo dT cellulose beads (NEB; S1408S), as described<sup>34</sup>. For some libraries, two sequential rounds of polyA selection were performed. Yeast RNA was fragmented in 10 mM ZnCl<sub>2</sub> at 94 °C for 5 min (total RNA) or 55 s (polyA+ RNA), and HeLa RNA was fragmented in 10 mM ZnAcetate at 60 °C for 10 min. Fragmented RNA was then precipitated.

CMC treatment of RNA fragments was as follows<sup>30</sup>. RNA was denatured in 5 mM EDTA at 80 °C for 2 min, and then placed on ice. 0.5 M CMC in BEU buffer (7 M urea, 4 mM EDTA, 50 mM bicine, pH 8.5) was added to a final concentration of 0.2 or 0.4 M CMC (4× RNA volume). CMC modification was carried out at 40 °C for 30 min, followed by ethanol precipitation. Subsequent reversal of modification of Us and Gs was carried out in NaCO<sub>3</sub> buffer (50 mM sodium-carbonate, pH 10.4, 2 mM EDTA) at 50 °C for 2 h, followed by precipitation. In parallel mock-treated samples were incubated in BEU buffer without CMC.

RNA fragments were dephosphorylated with CIP (NEB; M0290) and PNK (NEB; M0201), followed by size selection and elution of 80–100, 100–120, and 120–140-nt fragments on an 8% urea-TBE polyacrylamide gel electrophoresis (PAGE) gel, followed by precipitation. RNA fragments were eluted from gel slices overnight at 4 °C with gentle rocking in 400 µl RNA elution buffer (300 mM NaOAc pH 5.5, 1 mM EDTA, 100 U ml<sup>-1</sup> RNasin (Promega; N2615)). Ligation of a pre-adenylated 3' adaptor (IDT; /5Phos/TGGAATTCTCGGGTGCCAAGG/3ddC/) was carried out with T4 RNA ligase (NEB; M0204) in 1× buffer without ATP (50 mM Tris-HCl, pH 7.8, 10 mM MgCl<sub>2</sub>, 10 mM DTT) at 22 °C for 2.5 h, followed by precipitation.

Reverse transcription (RT) was carried out using AMV-RT (Promega; M5108) with the following conditions. The reverse transcription primer (IDT; /5Phos/GATCGTGGACTGTAGAACTCTGAACCTGTGCGTGGTCCGCGTATCATT/Sp18/CACTCA/Sp18/GCCTTGGCACCCGAGAATTCCA) and RNA were denatured and annealed in reverse transcription buffer (50 mM Tris-Cl pH 8.6, 60 mM NaCl, 10 mM DTT). After annealing, dNTPs (3.3 mM each final) and MgCl<sub>2</sub> (6 mM final) were added, and reverse transcription was carried out at 42 °C for 1 h. Truncated cDNAs were size-selected and purified on an 8% urea-TBE PAGE gel, followed by precipitation. cDNAs were eluted from gel slices overnight at room temperature with gentle rocking in 400 µl DNA elution buffer (300 mM NaCl, 10 mM Tris, pH 8.0).

cDNAs were circularized with circLigase (Epicentre; CL4115K), and amplified by PCR with Phusion (NEB; M0530) with the forward primer (IDT; AATGATACGGCAGCACCGA), and a barcoded reverse primer (IDT; CAAGCAGAAGACGGCATAACGAGATXXXXXXGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA). PCR products were gel-purified, precipitated and sequenced on an Illumina HiSeq 2000.

**Sequencing data analysis.** RNA-seq data was analysed with in-house Bash and Python scripts unless otherwise specified. For yeast libraries, adaptor sequences were trimmed using Cutadapt<sup>35</sup>, and were subsequently mapped to the *S. cerevisiae* genome downloaded from the Saccharomyces Genome Database (SGD) on 9/2/2011. Mapping to the genome and defined splice junctions (UCSC, sacCer3) was performed using Tophat2<sup>36</sup>. Multiply mapping reads were allowed. Using SAMtools to exclude multiply mapping reads affected  $\Psi$ s called in repetitive or paralogous features, but not  $\Psi$ s identified in other features<sup>37</sup>.

Trimmed reads from HeLa libraries were mapped with Bowtie1 allowing up to 2 mismatches to a database of spliced transcripts (hg19 sequence, and transcripts downloaded from UCSC on 1 August 2012) containing the transcript with the longest coding sequence, or the longest transcript for non-coding genes<sup>38</sup>. Multiply mapping

reads were allowed for generation of receiver operating characteristic curves and MetaPsi plots, but were excluded for  $\Psi$  predictions.

**Identification of  $\Psi$ .** The yeast transcriptome (downloaded from SGD on 9 February 2011) with annotated 5' and 3' UTRs was used to identify new sites of pseudouridylation<sup>39</sup>. Where annotated 5' and 3' UTRs were not available, median UTR lengths were used. To identify new sites of pseudouridylation in HeLa cells the human transcriptome described above was used. For a given +/–CMC library pair, the –CMC libraries were first scaled to the size of the +CMC libraries. Peak values were calculated for each position 1 nt 3' of a U (peak position) in all features with an average per nucleotide read coverage greater than a specified read cutoff:

$$\text{peak}^+ = \text{ws} \times \frac{r^+ - r^-}{\text{wr}^+ + \text{wr}^-}$$

Where  $r^+$  and  $r^-$  indicate the number of reads whose 5' ends map to the position being examined in the +CMC and –CMC libraries, respectively,  $\text{wr}^+$  and  $\text{wr}^-$  represent the numbers of reads whose 5' ends map to a window centred at the position being examined (exclusive of reads at that position), and  $\text{ws}$  specifies the size of this window (exclusive of that position). Sites with peak positions greater than a specified peak cutoff were flagged as potential  $\Psi$ . To filter out false positives reproducibility of peak calling in a certain number of libraries was required.

Window size ( $\text{ws}$ ) was set to 150 for all analyses. Only features surpassing an average reads/nt threshold (read cutoff) were considered. For high density yeast, the read cutoff was set to 0.0, the peak cutoff was set to 1.0, and reproducibility was required in 10 of 14 libraries. For log phase yeast, the peak values from log phase data were calculated for the high density identified  $\Psi$ . For both serum fed and serum starved HeLa cells, the read cutoff was set to 1.0, the peak cutoff was set to 2.0, and reproducibility was required in 4 of 4 libraries. A subset of called  $\Psi$  in HeLa cells came from very narrow (<20 nt) regions of uniquely mapping reads. These calls were considered unreliable and were removed.

**MetaPsi plots and ROC curves.** For a given  $\Psi$ , the reads at each position in a 51-nt window centred at the  $\Psi$  were normalized to the average reads per nucleotide within the window. These windows of normalized reads were then averaged for all known  $\Psi$  in yeast rRNA and U2 or human rRNA and snRNA, yielding a metaPsi. Given the close spacing of  $\Psi$  in these features, the number of  $\Psi$  at each position in the metaPsi window was also plotted.

To generate receiver operating characteristic (ROC) curves for a given +/–CMC library pair, the –CMC libraries were first scaled to the size of the +CMC libraries, and peak values were calculated (see above) for each position 1 nt 3' of a U or  $\Psi$  in the features above. Additionally, a –CMC peak value was determined:

$$\text{peak}^- = \text{ws} \times \frac{r^-}{\text{wr}^+ + \text{wr}^-}$$

Parameters are as defined earlier. A range of 10,000 equally spaced cutoff scores were chosen spanning the range of observed peak values. At each cutoff score, the true positive and false positive rates were calculated, and plotted.

**Estimation of false discovery rate.** The lower bound for FDR was estimated from the observed FDR for the rRNA. The upper bound of FDR for lowly expressed genes was estimated by randomly down-sampling reads in the rRNA and U2 snRNA to a level of coverage comparable to that of lowly expressed mRNAs with  $\Psi$ s. These randomizations were performed 14 times followed by  $\Psi$  calling on down-sampled libraries as described above. This number should be considered a rough estimate because the ribosomal RNA may not provide a perfect basis for estimating the FDR of  $\Psi$  calling in mRNA. The observed number of false positives in the rRNA and snRNA under the criteria used to call  $\Psi$ s in mRNA was two incorrect  $\Psi$  calls in 1905 U residues (0.1%).

**Defining  $\Psi$  regulation and factor-dependence.** To determine if a given  $\Psi$  was condition dependent the median peak values between two conditions were compared. For yeast only wild-type peak values were included. A twofold or greater change between conditions was considered regulated.

$\Psi$ s were identified as Pus-dependent if the peak heights in both biological replicates of a given *pusA* strain were less than 25% of the median peak height for that  $\Psi$  across all libraries. For a given  $\Psi$  to be considered PUS-dependent, we required at least one replicate for a given *pusA* strain to have sufficient reads in the 150-nt window surrounding the  $\Psi$  to be greater than 25% of the median reads in that window for all libraries.

**snoRNA target site predictions and analysis.** To identify potential sites of pseudouridylation within yeast mRNAs, the yeast transcriptome (described above) was scanned for sites that perfectly match the known target sequences of all yeast Box H/ACA snoRNAs, allowing mismatches at bases that are unpaired in known target sites<sup>40</sup>.

For the analysis presented in Extended Data Fig. 7a, b, 10,000 randomizations were performed. For each trial, a random U was chosen for each non-repetitive computationally predicted snoRNA target, and was matched to the same gene, and

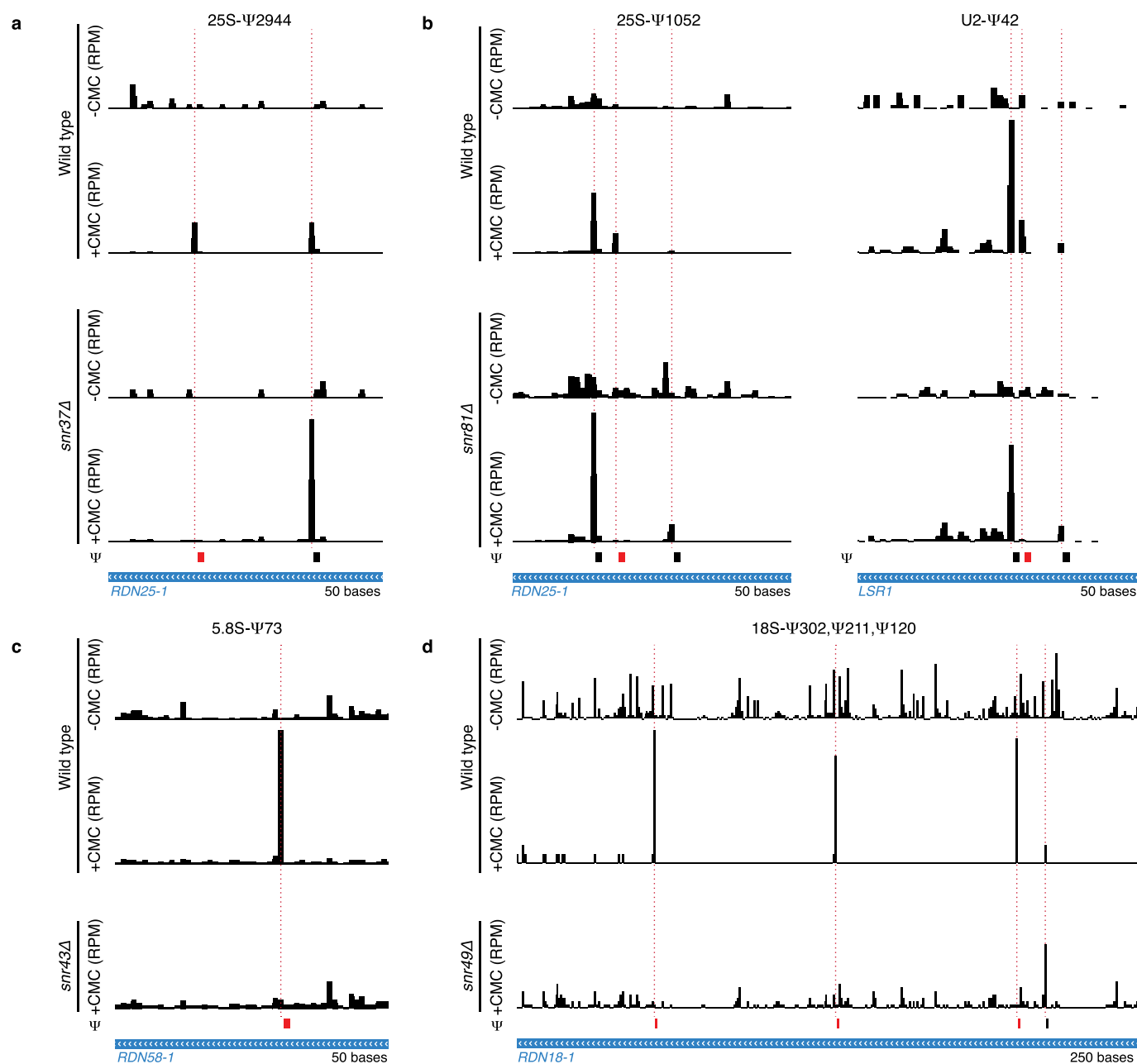


transcript feature (5' UTR, CDS, 3' UTR). Each randomized set of Us was used to generate a metaPsi from the pooled reads of four libraries, and the differences in mean normalized reads between the +CMC and -CMC pools at the peak position were calculated. The distributions of these values were plotted in a histogram, and compared to the values for the computationally predicted snoRNA target sites.

**Plotting and other analyses.** Sequences and structures of tRNAs were obtained from tRNAdb, and  $\Psi$  locations were previously published<sup>41–43</sup>. Motifs were generated using WebLogo 3.3 using default settings, and the modified position was changed to a  $\Psi$  after logo generation<sup>41</sup>. UCSC genome browser was used to generate plots of rpm data, and matplotlib was used to generate the remainder of graphs<sup>44</sup>. RPKMs (reads per kilobase of exon sequence per million exon reads) were calculated from -CMC libraries. GO analysis was performed using the YeastMine feature of SGD (<http://yeastmine.yeastgenome.org/>). The set of genes whose coverage was sufficient to reproducibly call  $\Psi$ s was used as a background set. These 5,278 genes had average RPKMs in post-diauxic wild type cultures of  $\geq 9.25$ , the level of expression of the lowest expressed pseudouridylated mRNA called by our algorithm.

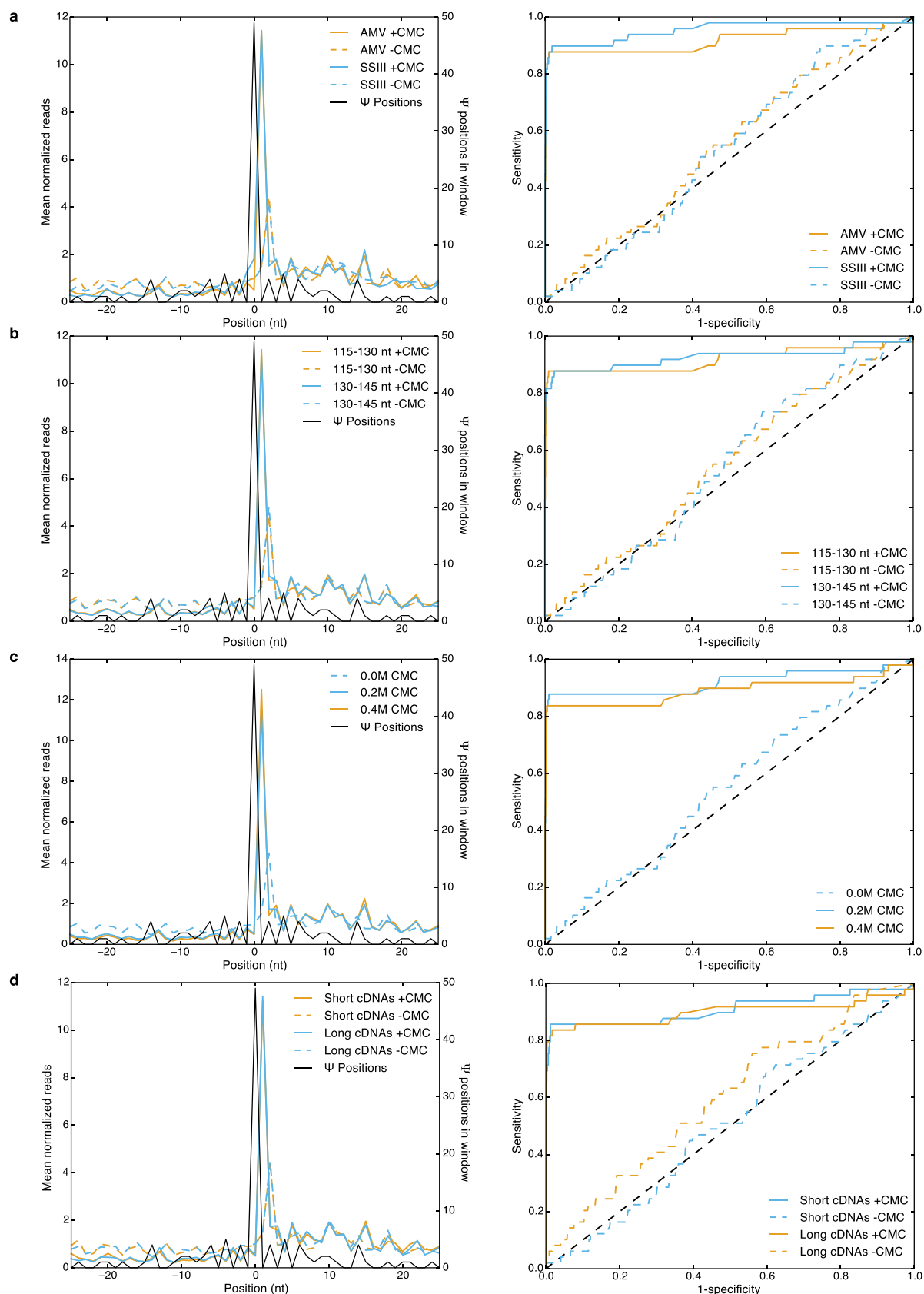
31. Longtine, M. S. *et al.* Additional modules for versatile and economical PCR-based gene deletion and modification in *Saccharomyces cerevisiae*. *Yeast* **14**, 953–961 (1998).
32. Winzeler, E. A. *et al.* Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* **285**, 901–906 (1999).
33. Collart, M. A. & Oliviero, S. Preparation of yeast RNA. *Curr. Prot. Mol. Biol.* **Ch. 13**, Unit-13.12 (2001).
34. Sambrook, J. & Russell, D. W. *Molecular Cloning* (Cold Spring Harbor Laboratory Press, 2001).
35. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnetjournal* **17**, 10–12 (2011).
36. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).
37. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
38. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
39. Xu, Z. *et al.* Bidirectional promoters generate pervasive transcription in yeast. *Nature* **457**, 1033–1037 (2009).
40. Piekna-Przybylska, D., Decatur, W. A. & Fournier, M. J. New bioinformatic tools for analysis of nucleotide modifications in eukaryotic rRNA. *RNA* **13**, 305–312 (2007).
41. Crooks, G. E., Hon, G., Chandonia, J.-M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res.* **14**, 1188–1190 (2004).
42. Darty, K., Denise, A. & Ponty, Y. VARNA: interactive drawing and editing of the RNA secondary structure. *Bioinformatics* **25**, 1974–1975 (2009).
43. Jühling, F. *et al.* tRNAdb 2009: compilation of tRNA sequences and tRNA genes. *Nucleic Acids Res.* **37**, D159–D162 (2009).
44. Hunter, J. D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).





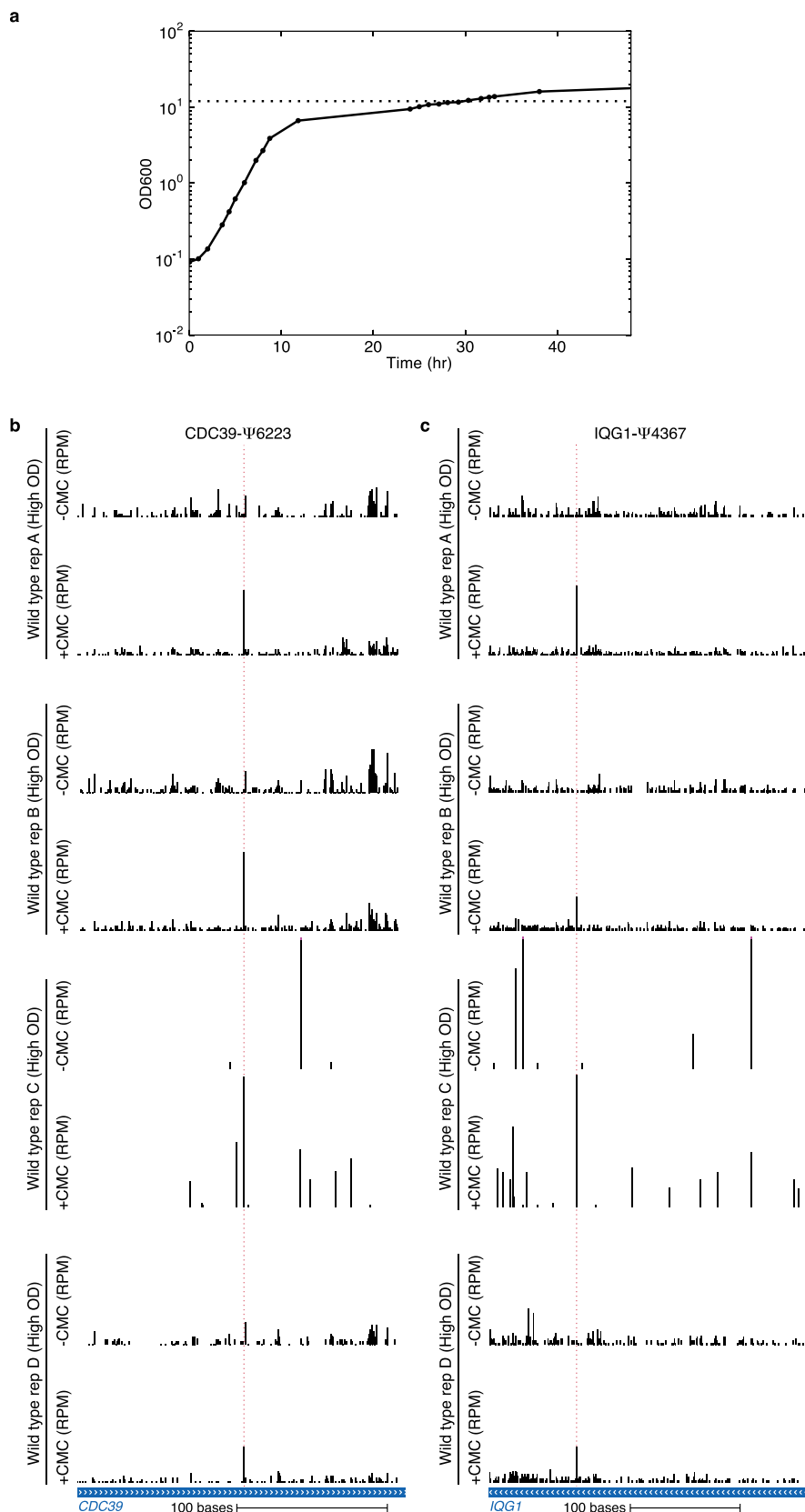
**Extended Data Figure 1 | Detection of specific snoRNA target sites by Pseudo-seq.** Pseudo-seq was performed on wild-type ( $n = 4$ ), *snr37Δ* ( $n = 2$ ), *snr81Δ* ( $n = 2$ ), *snr43Δ* ( $n = 2$ ) and *snr49Δ* ( $n = 2$ ) yeast strains. Cultures were harvested at high density (a, b) or log phase (c, d). Ψs dependent on the deleted snoRNA are indicated in red. CMC-dependent peaks of reads are indicated with dashed red lines. Traces are representative of indicated number of biological replicates. a, Pseudo-seq reads in *RDN25-1* (chrXII: 452221–452270) showing *SNR37*-dependence of 25S-Ψ2944. b, Pseudo-seq reads in *RDN25-1*

(chrXII: 454111–454160, left), and U2 snRNA (*LSR1*, chrII: 681791–681840, right) showing *SNR81*-dependence of 25S-Ψ1052 and U2-Ψ42. c, Pseudo-seq reads in *RDN58-1* (chrXII: 455466–455515) showing *SNR43*-dependence of 5.8S-Ψ73. *SNR43*-dependent 25S-Ψ960 was not consistently detected in wild type owing to an overlapping CMC-independent reverse transcription stop. d, Pseudo-seq reads in *RDN18-1* (chrXII: 457361–457610) showing *SNR49*-dependence of 18S-Ψ302, 18S-Ψ211, and 18S-Ψ120. 25S-Ψ990 was also detected as *SNR49*-dependent (data not shown).



**Extended Data Figure 2 | Technical variations of Pseudo-seq give similar results.** **a–d**, MetaPsi plots (left), and ROC curves (right) for various library prep conditions  $n = 1$  for each condition. CMC-treated samples (solid) and mock-treated samples (dashed) are indicated. **a**, Comparison of AMV-RT (orange) and SuperScript III (blue) (0.2 M CMC; 115–130 nt, 100–115 nt

fragments respectively). **b**, Comparison of 115–130 nt (orange), and 130–145 nt (blue) RNA fragment sizes (AMV-RT; 0.2 M CMC). **c**, Comparison of 0.2 M CMC (blue), and 0.4 M CMC (orange) (AMV-RT; 115–130 nt RNA). **d**, Comparison of shorter (orange) and longer (blue) truncated reverse transcription fragment sizes (AMV-RT; 115–130 nt RNA; 0.2 M CMC).

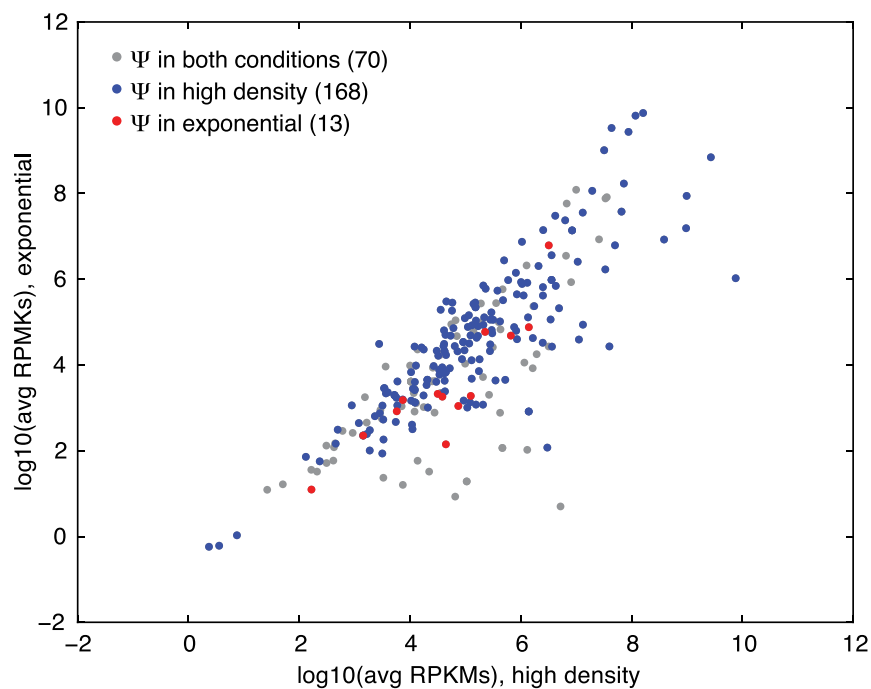


**Extended Data Figure 3 | Identification of pseudouridines in lowly expressed genes using multiple replicates.** **a**, Growth curves for wild-type yeast were grown in YPD. An  $A_{600\text{ nm}}$  of 12 is indicated by the horizontal dotted line. **b**, **c**, Pseudo-seq was performed on polyA<sup>+</sup> RNA isolated from high-density wild-type yeast strains. CMC-dependent peaks of reads are indicated

with dashed red lines. **b**, Pseudo-seq reads from  $n = 4$  biological replicates in **a** *CDC39* (chrIII: 286226–286445, 12.3 average RPKMs), and **c**, *IQG1* (chrXVI: 90655–90955, 12.4 average RPKMs) showing *CDC39*-Ψ6223 and *IQG1*-Ψ4367, respectively.

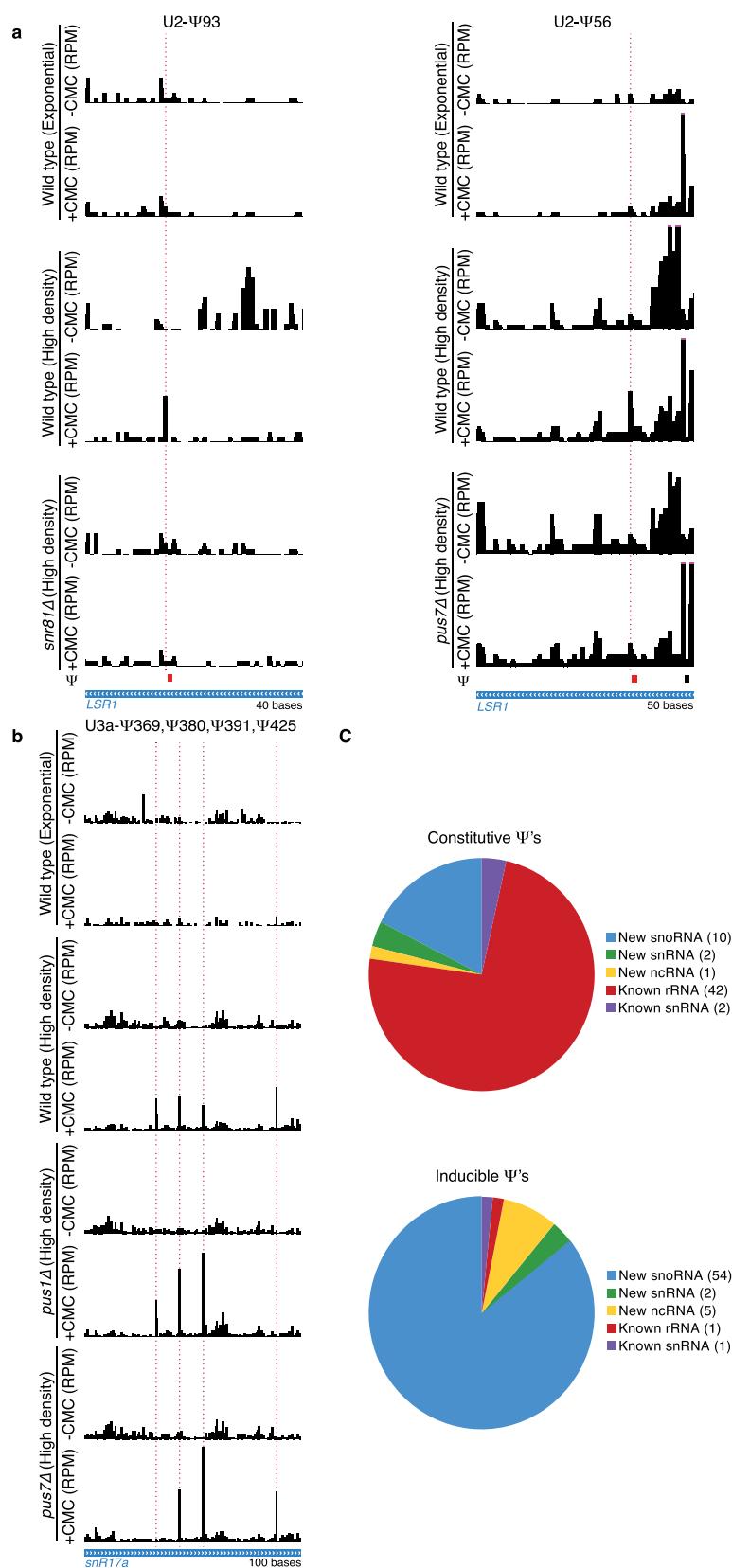






**Extended Data Figure 5 | Expression levels minimally affect identification of yeast mRNAs displaying regulated pseudouridylation.** A plot of log-transformed average RPKMs in high-density versus log-phase yeast for all

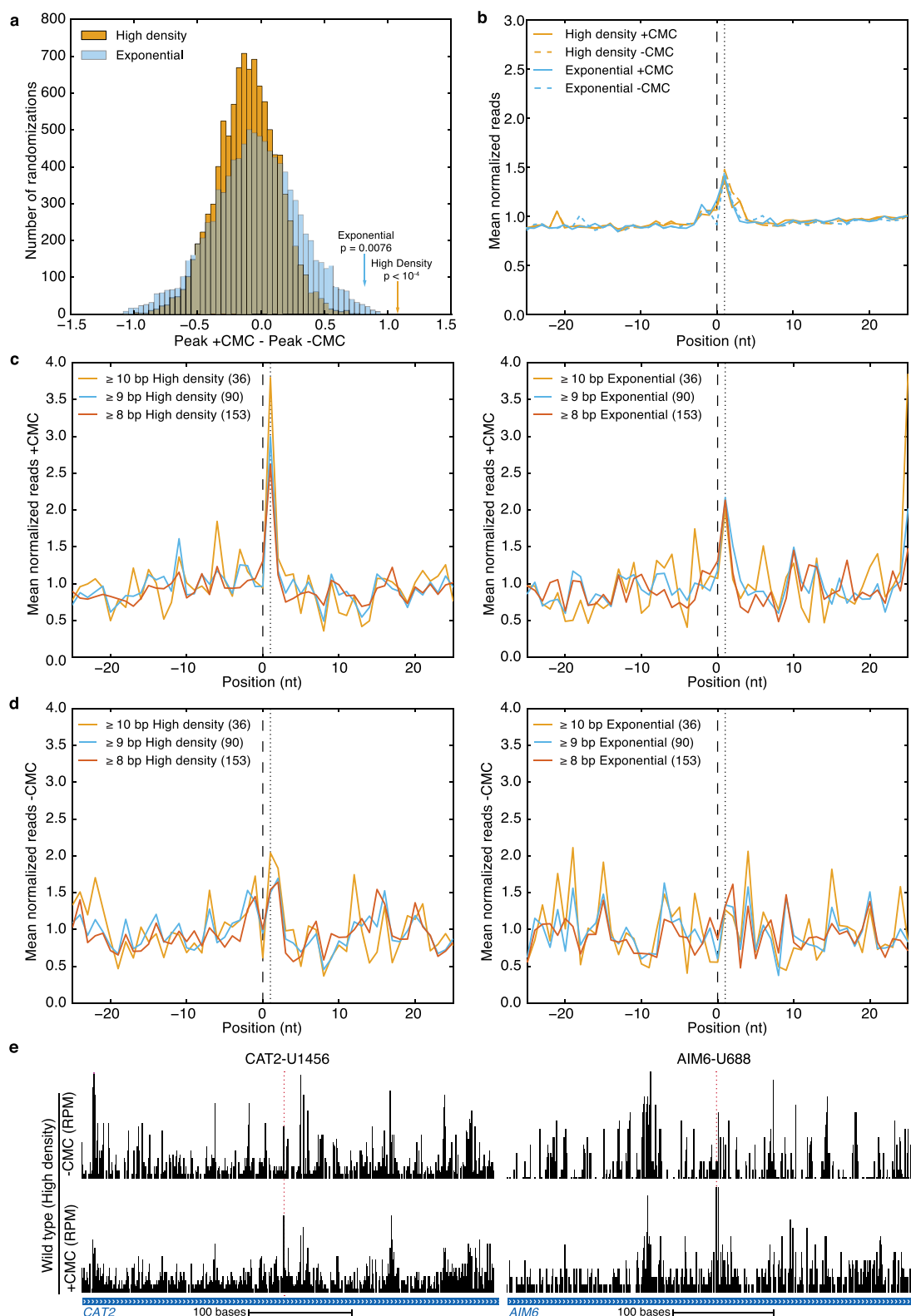
coding genes with a  $\Psi$  identified by Pseudo-seq  $n = 4$  biological replicates for each condition. All genes (grey), genes with a high density induced  $\Psi$  (blue), and genes with a log phase induced  $\Psi$  (red) are indicated.



### Extended Data Figure 6 | Inducible pseudouridylation of ncRNAs.

**a, b**, Pseudo-seq was performed on wild-type ( $n = 4$ ), *snr81Δ* ( $n = 2$ ), *pus1Δ* ( $n = 2$ ) and *pus7Δ* ( $n = 2$ ) yeast strains grown to high density. CMC dependent peaks of reads are indicated with a dashed red line. Traces are representative indicated number of biological replicates. **a**, Pseudo-seq reads in U2 snRNA (*LSR1*; chrII: 681751–681790, left; chrII: 681769–681818, right) showing

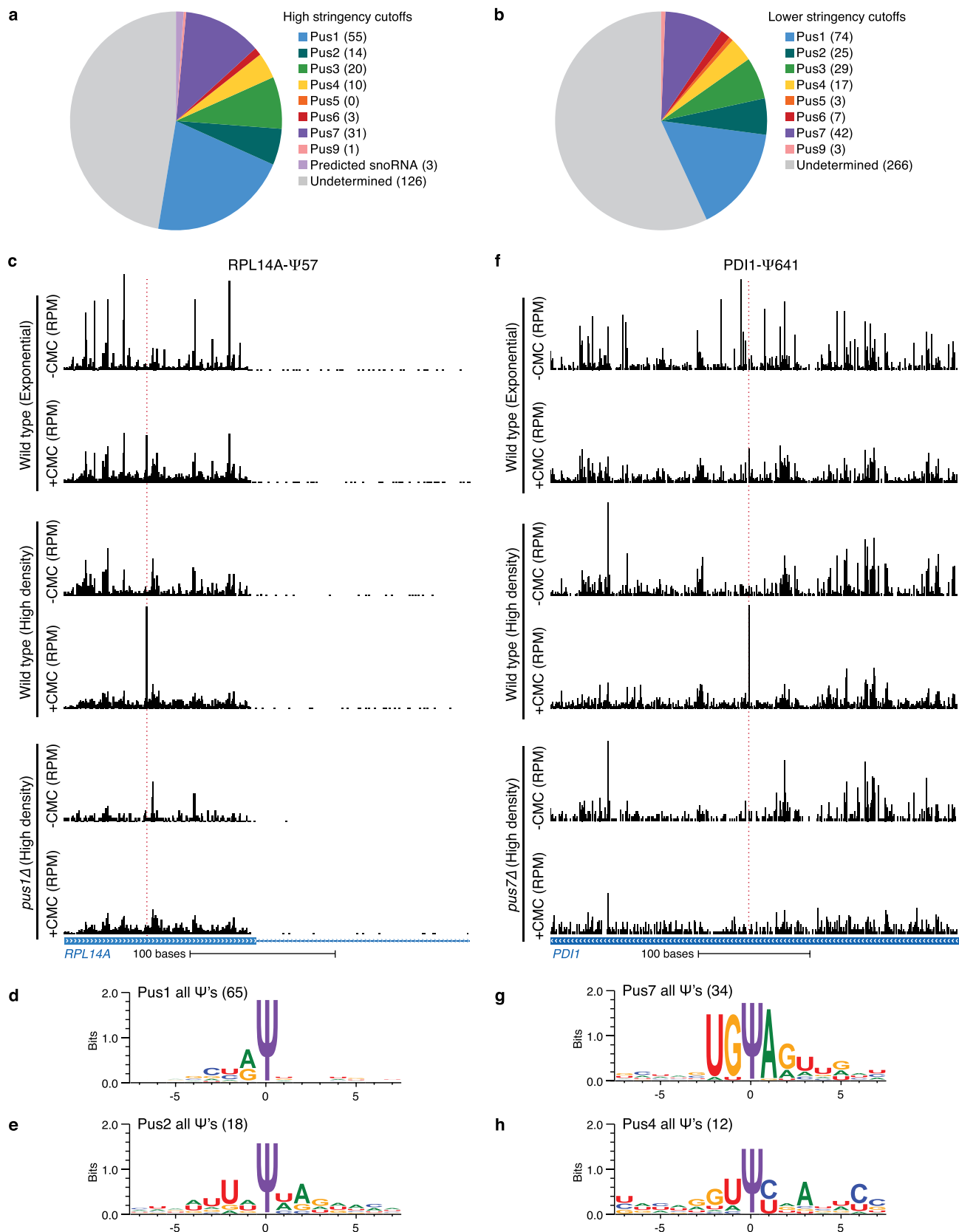
*SNR81*-dependence of U2-Ψ93, and *PUS1*-dependence of U2-Ψ56. Both are dependent on growth to high density. **b**, Pseudo-seq reads in U3a snoRNA (*SNR17A*, chrXV: 780461–780560) showing *snr17a*-Ψ369 (*PUS7*-dependent), *snr17a*-Ψ380, *snr17a*-Ψ391 and *snr17a*-Ψ425 (*PUS1*-dependent). **c**, Summaries of the numbers of Ψs called in ncRNAs by Pseudo-seq. Indicated are constitutive Ψs (top), and inducible Ψs (bottom).



### Extended Data Figure 7 | Analysis of potential snoRNA targets.

**a–d**, Pseudo-seq was performed on wild-type yeast in log phase, or grown to high density. Reads from  $n = 4$  biological replicate libraries for each condition were pooled. **b–d**, Indicated are the predicted snoRNA target site (black, dashed), and the expected peak of CMC-dependent reads (black, dotted). **a, b**, Results of analysis on sets of random Us. **a**, A histogram of the differences (+CMC – –CMC) in mean normalized reads at the +1 peak position for 10,000 randomizations for high density (orange) and log phase (blue). The normalized read values for the computationally predicted  $\Psi$ s in exponential

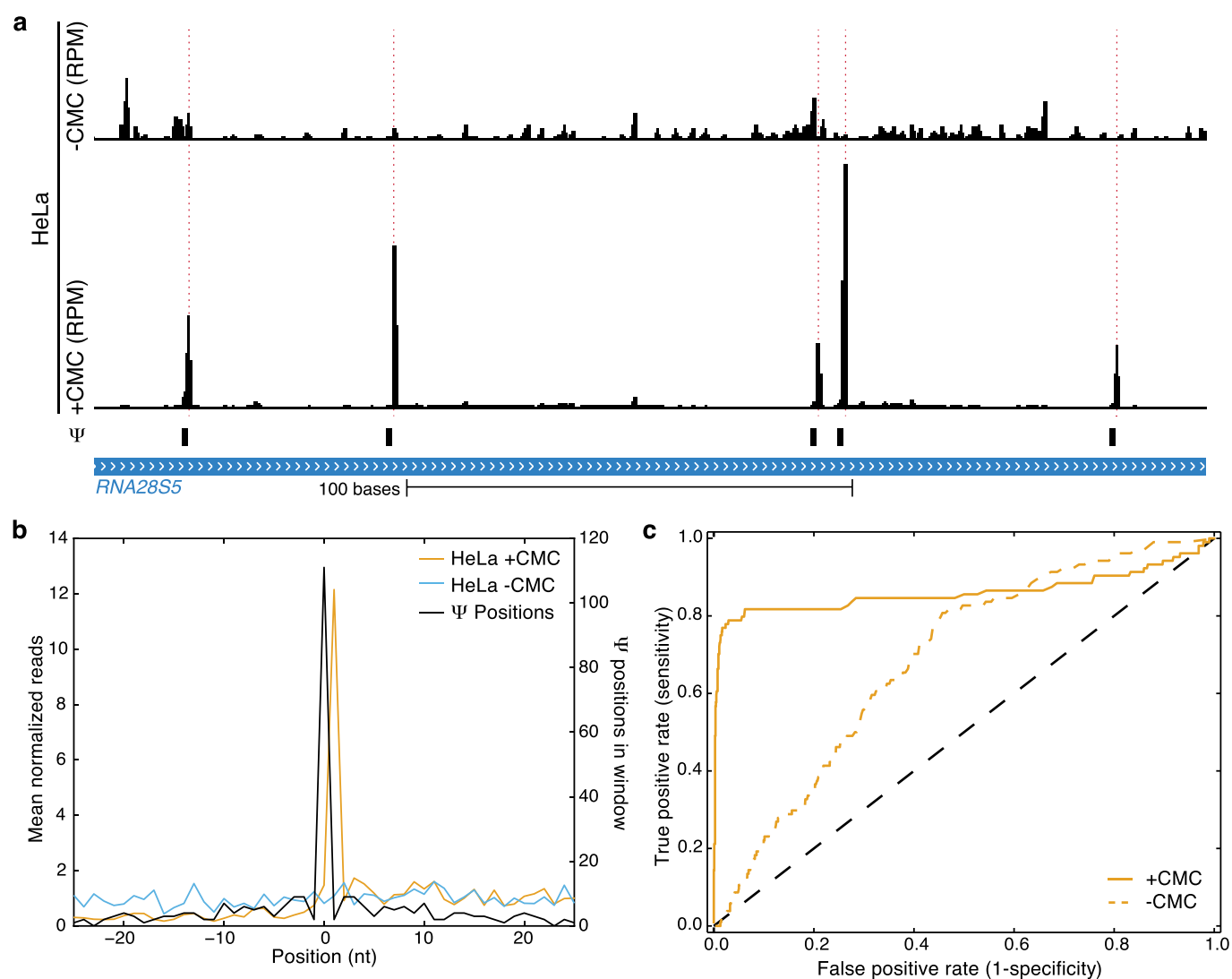
and high density samples are indicated by arrows. **b**, An averaged metaPsi plot for all randomizations. **c, d**, +CMC (**c**) and –CMC (**d**) MetaPsi plots for computationally predicted  $\Psi$ s separated by base pairing. Sites with 8 or more (red), 9 or more (blue), and 10 or more (orange) base pairs are indicated. Data for high density (left), and log phase (right) are indicated. **e**, Pseudo-seq reads for computationally predicted  $\Psi$ s, CAT2 (chrXII: 193995–19450, left), and AIM6 (chrIV: 31135–31550, right). Traces are representative of six biological replicates.



**Extended Data Figure 8 | Mechanisms of Pus-dependent pseudouridylation.** **a, b**, Summaries of the *PUS*-dependence of called Ψ's using higher stringency cut-offs (10/14 libraries) (**a**) and lower stringency cut-offs (9/14 libraries) (**b**). **c, f**, CMC-dependent peaks of reads are indicated with dashed red lines. Traces are representative of  $n = 4$  (wild type), and  $n = 2$  (*pus1Δ*)

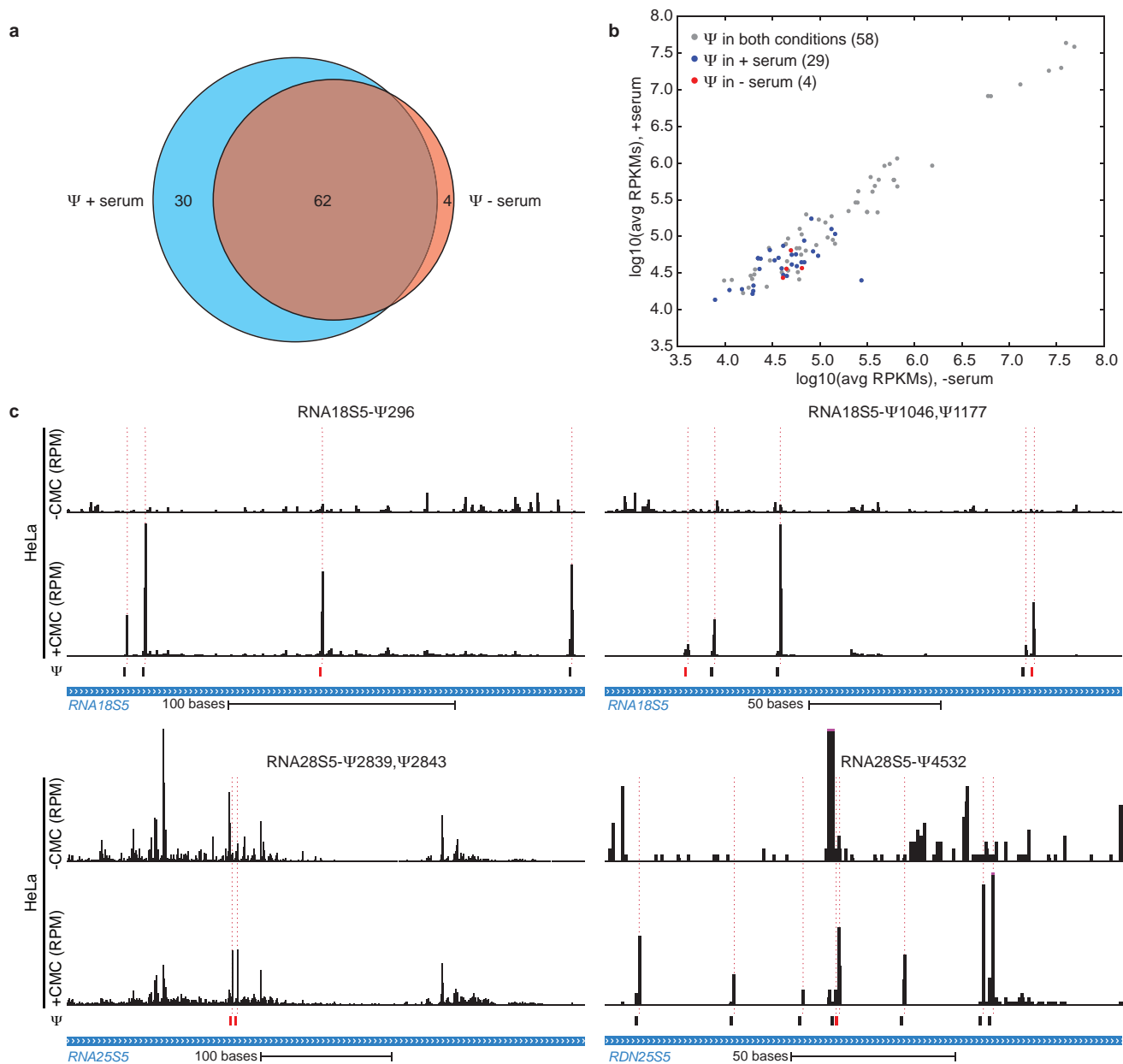
biological replicates. Pseudo-seq reads for *RPL14A* (**a**, chrXI: 431901–432200) and *PDI1* (**d**, chrIII: 49401–48760) showing *PUS1*- and *PUS7*-dependency, respectively. Both are dependent on growth to high density. **d, e, g, h**, WebLogo 3.3 was used to generate motifs for *PUS1* (**d**), *PUS2* (**e**), *PUS7* (**g**) and *PUS4* (**h**).





**Extended Data Figure 9 | Positive controls for human RNA Pseudo-seq.** **a**, Pseudo-seq reads for *RDN28S5* (1516–1765) containing five known  $\Psi$ s (28S- $\Psi$ 1536, 28S- $\Psi$ 1582, 28S- $\Psi$ 1677, 28S- $\Psi$ 1683 and 28S- $\Psi$ 1744). CMC-dependent peaks of reads are indicated with dashed red lines. Traces are representative of  $n = 5$  biological replicates. **b**, A metaPsi plot of mean

normalized reads (left axis) for +CMC libraries (orange), and -CMC libraries (blue). The number of  $\Psi$ s at each position in the metaPsi window is indicated (black, right axis). **c**, A ROC curve of the Pseudo-seq signal for all known  $\Psi$ s in rRNA and snRNA.



**Extended Data Figure 10 | New pseudouridines in human RNAs.** **a**, A Venn diagram showing the overlap of mRNA pseudouridylation events between serum-fed and serum-starved HeLa cells. **b**, A plot of log-transformed average RPKMs in serum-starved versus serum-fed HeLa for all coding genes with a Ψ identified by Pseudo-seq. All genes with a Ψ (grey), genes with a Ψ induced in plus serum cells (blue), and genes with a Ψ induced in serum-starved cells

(red) are indicated. **c**, Pseudo-seq reads for *RDN18S5* (184–411) (top, left), *RDN18S5* (1015–1210) (top, right), *RDN28S5* (2713–3108) (bottom, left), and *RDN28S5* (4461–4618) (bottom, right). CMC-dependent peaks of reads are indicated with dashed red lines, and highlighted Ψs are indicated by red boxes. Traces are representative of  $n = 4$  biological replicates.

# Crystal structure of the RNA-guided immune surveillance Cascade complex in *Escherichia coli*

Hongtu Zhao<sup>1,2</sup>, Gang Sheng<sup>1</sup>, Jiuyu Wang<sup>1</sup>, Min Wang<sup>1</sup>, Gabor Bunkoczi<sup>3</sup>, Weimin Gong<sup>1</sup>, Zhiyi Wei<sup>4</sup> & Yanli Wang<sup>1</sup>

Clustered regularly interspaced short palindromic repeats (CRISPR) together with CRISPR-associated (Cas) proteins form the CRISPR/Cas system to defend against foreign nucleic acids of bacterial and archaeal origin<sup>1–9</sup>. In the I–E subtype CRISPR/Cas system, eleven subunits from five Cas proteins (CasA<sub>1</sub>B<sub>2</sub>C<sub>6</sub>D<sub>1</sub>E<sub>1</sub>) assemble along a CRISPR RNA (crRNA) to form the Cascade complex<sup>10–13</sup>. Here we report on the 3.05 Å crystal structure of the 405-kilodalton *Escherichia coli* Cascade complex that provides molecular details beyond those available from earlier lower-resolution cryo-electron microscopy structures. The bound 61-nucleotide crRNA spans the entire 11-protein subunit-containing complex, where it interacts with all six CasC subunits (named CasC1–6), with its 5′ and 3′ terminal repeats anchored by CasD and CasE, respectively. The crRNA spacer region is positioned along a continuous groove on the concave surface generated by the aligned CasC1–6 subunits. The five long β-hairpins that project from individual CasC2–6 subunits extend across the crRNA, with each β-hairpin inserting into the gap between the last stacked base and its adjacent splayed counterpart, and positioned within the groove of the preceding CasC subunit. Therefore, instead of continuously stacking, the crRNA spacer region is divided into five equal fragments, with each fragment containing five stacked bases flanked by one flipped-out base. Each of those crRNA spacer fragments interacts with CasC in a similar fashion. Furthermore, our structure explains why the seed sequence, with its outward-directed bases, has a critical role in target DNA recognition. In conclusion, our structure of the Cascade complex provides novel molecular details of protein–protein and protein–RNA alignments and interactions required for generation of a complex mediating RNA-guided immune surveillance.

Our 3.05 Å crystal structure of the *E. coli* Cascade complex exhibits a sea-horse-shaped architecture, similar to previous cryo-electron microscopy structures<sup>10</sup>. Instead of forming a simple, compact structure, the 11 subunits are assembled into two structural layers. Six CasC proteins, together with CasD and CasE, tightly pack to form the outer layer (Fig. 1b–d and Extended Data Fig. 1). Within this outer arch-like configuration, the six CasC proteins labelled C1 to C6 are assembled into a symmetry-related helical alignment defined by a ~40° rotation, together with a vertical shift of 25 Å, per subunit. CasE and CasA cap the two ends of the CasC helical arrangement by associating with CasC1 and CasC6, respectively. The inner layer, consisting of the CasA and the CasB dimer, is connected to the outer layer mainly via CasA–CasD interactions as well as multiple relatively weak CasC contact sites at their lateral regions (Fig. 1c). We observed all 61 nucleotides of the crRNA, with the 32-nucleotide spacer shown in red and the two flanking repeat segments (also called 5′- and 3′-handle) shown in orange in Fig. 1. The crRNA is properly positioned along the long outer layer (Extended Data Fig. 2). Interestingly, each subunit in the outer layer contains a long β-hairpin protruding from its structured core (Fig. 1d).

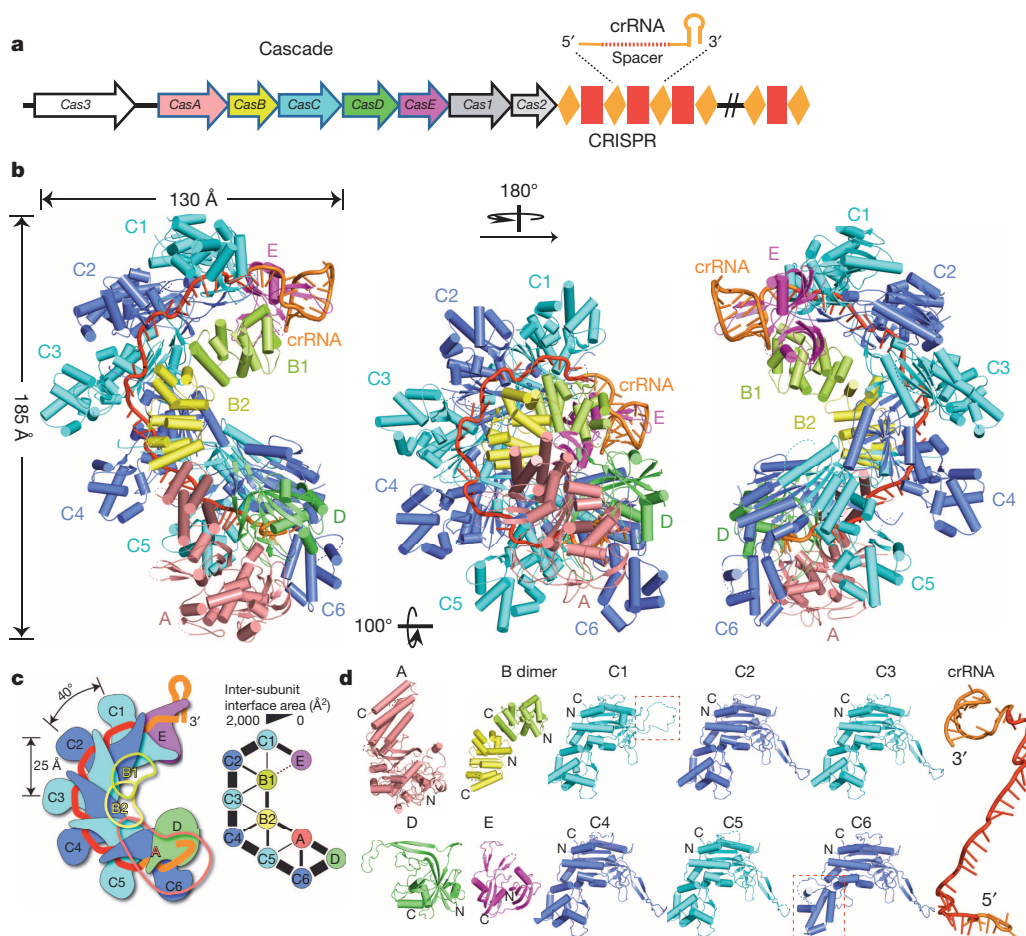
The CasE endonuclease subunit at the head end of the Cascade complex adopts a two-ferredoxin fold, similar to the RNA-bound structures of *Thermus thermophilus* Cse3<sup>14,15</sup> and *Pseudomonas aeruginosa* Csy4<sup>16</sup>.

The 3′-end of the crRNA adopts a stem loop-fold, with the RNA stem anchored within the positively charged cleft of the carboxy-terminal domain of CasE (Fig. 2a, b and Extended Data Fig. 3a). Its β6–β7 hairpin, a universal structural feature of the CasE family, contains several positively charged residues that contact the crRNA backbone. The 3′-end repeat of crRNA is recognized by the side chains of R111<sup>CasE</sup> and R113<sup>CasE</sup>, which insert into the major groove of the RNA stem and contact the bases of G(+10) and G(+19) (Extended Data Fig. 3b, c). With the following continuous turn, the 3′ terminal repeat surrounds the β6–β7 hairpin, a feature not observed previously, and is probably responsible for proper anchoring of the crRNA in the Cascade complex (Fig. 2b). At the opposite face of the crRNA stem-binding surface, a V-shaped cleft is formed between the two domains of CasE. There, a specific loop from CasC1 residues 197–208 inserts into this V-shape cleft of CasE via hydrophobic interactions (Fig. 2c and Extended Data Fig. 3d).

Once the crRNA is Cascade-bound, the 8-nucleotide repeat of its 5′-end takes on a hook-like shape, with a sharp turn at the G1–G(–1) and G(–1)–C(–2) steps, preceded by continuous stacking within the C(–3)–A(–5) steps, and three residues pointing into a different direction at each base (Fig. 2d). The 5′-handle is buried into a pocket formed by CasD, CasC6 and CasA in a sequence-specific manner (Fig. 2d–g). Residues C(–2), C(–3), A(–5) and U(–7) interact with CasD and CasC6 sequence specifically. The base of G(–1) stacks onto the side-chain of L89<sup>CasD</sup>. R108<sup>CasD</sup>, which is located within the β5–β6 hairpin, forms hydrogen bonds with the base of C(–2) (Fig. 2e). The CRISPR/Cas system cannot launch a defence against the invading plasmid once R108<sup>CasD</sup> is substituted by Ala (Extended Data Fig. 4a). F129<sup>CasA</sup> and K177<sup>CasC6</sup> form respective hydrogen bonds with the bases of C(–3) and A(–5), which are further stabilized by stacking with the side chains of F129<sup>CasA</sup> and R206<sup>CasD</sup>, respectively (Fig. 2e and f). The 5′-terminal residues A(–6), U(–7) and A(–8) are sandwiched by amino acid residues; that is, A(–6) between F208<sup>CasD</sup> and P19<sup>CasD</sup>, U(–7) between Y145<sup>CasD</sup> and the peptide plane of residues 38–39<sup>CasD</sup> on the α1-helix, and A(–8) between Y142<sup>CasD</sup> and D179<sup>CasC6</sup>–R48<sup>CasD</sup> pair. Additionally, the 2′ hydroxyl and phosphate groups of the 5′-handle provide hydrogen bonds and ionic bonds with CasD (Fig. 2f, g and Extended Data Fig. 2).

A specific β5–β6 hairpin of CasD, which contacts CasA and CasC6, has a critical role in 5′-handle binding. This hairpin extends across the RNA backbone to CasC6 through the gap between G(–1) and C(–2), in similar fashion to that of CasC2–6 (discussed below). In addition, a loop formed by residues 120–135<sup>CasA</sup> inserts into the cavity between the β-hairpin and the main body of CasD (Fig. 2h and Extended Data Fig. 4c), and a loop of CasC6 (residues 164–179) inserts into the opposite groove (Extended Data Fig. 4d), indicating a role of the β-hairpin for proper assembly of CasA and CasC6 into the Cascade complex. In the crRNA-free structure, this arm is highly flexible<sup>17</sup>, suggesting mutual stabilization between CasD and bound crRNA, which is supported by the fact that upon replacement of the β-hairpin by a non-functional (GGG)<sub>4</sub> linker, the Cascade complex cannot be assembled properly and Cascade-mediated immunity was abolished (Extended Data Fig. 4a, b, e, f).

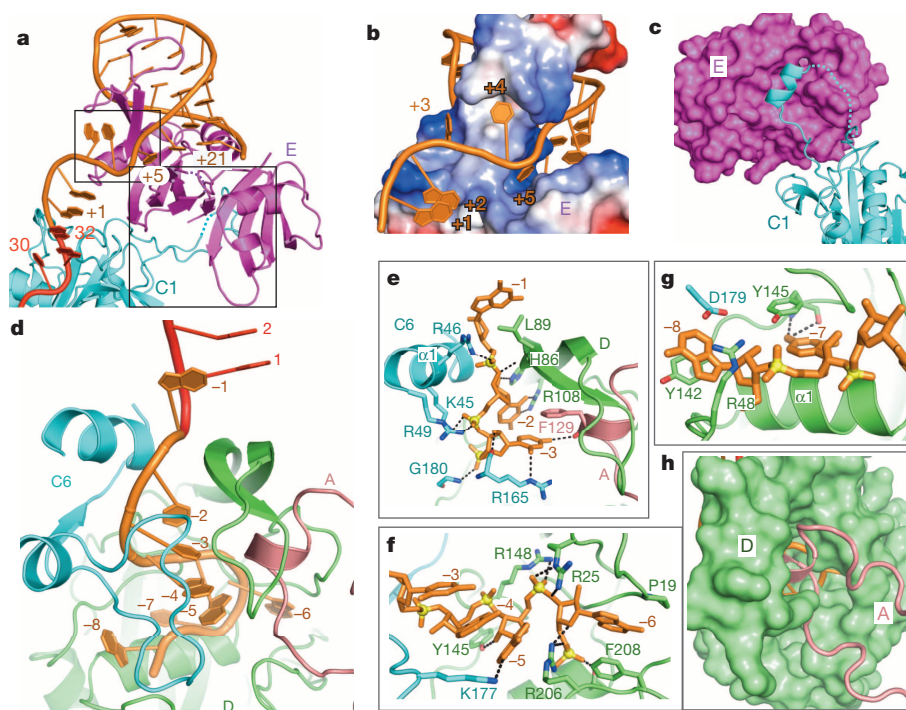
<sup>1</sup>Laboratory of RNA Biology, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China. <sup>2</sup>University of Chinese Academy of Sciences, Beijing 100049, China. <sup>3</sup>Cambridge Institute for Medical Research, Addenbrooke's Hospital, Hills Road, Cambridge CB2 0XY, UK. <sup>4</sup>Department of Biology, South University of Science and Technology of China, Shenzhen, 518055, China.



**Figure 1 | Crystal structure of the Cascade complex from *E. coli*.**  
**a**, The I–E subtype CRISPR system in *E. coli* (K12) consists of eight Cas proteins and CRISPR locus.  
**b**, Overall structure of the Cascade complex showing the structural uniqueness in both the crRNA recognition and the subunit organization. All subunits in the outer layer contain a long hairpin protruding from the structural core for stapling the crRNA spacer, for specific binding of the stem-loop structure, and for the outer layer organization. The sophisticated protein interaction network in the Cascade organization is shown.  
**d**, Individual components are shown.

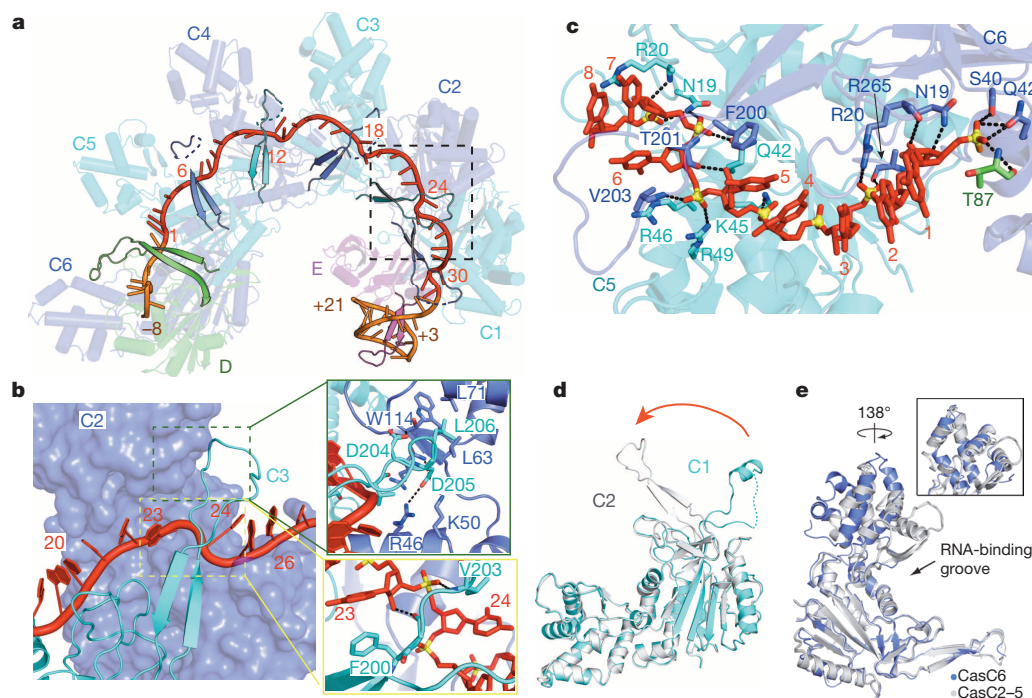
Six CasC subunits form the backbone of the Cascade complex. The proximal (inner side of the helix) and distal domains (outer side) of each CasC adopt a concave shape (Fig. 3). Except for CasC1, all CasC proximal domains contain long  $\beta$ -hairpins (Fig. 1d), which are flexible in the RNA-free structures<sup>18,19</sup>, and stabilized upon crRNA–Cascade assembly.

The  $\beta$ -hairpin of each CasC extends up into the groove of its preceding CasC subunit, where it is locked by interactions of the  $\beta$ -hairpin residues D204<sup>CasC</sup>, D205<sup>CasC</sup> and L206<sup>CasC</sup> with the surrounding groove residues (Fig. 3a, b and Extended Data Fig. 5a–c). Consistently, mutating D204–L206<sup>CasC</sup> to Ala abolishes Cascade-mediated immunity



**Figure 2 | The two ends of the crRNA are anchored.**  
**a**, The 3'-repeat segment winds the  $\beta$ 6– $\beta$ 7 hairpin of CasE. **b**, Positioning of non-helical residues (+1)–(+5). U(+5) is inserted into the cleft between two hairpins. U(+4) and G(+3) is inserted into the concave surface. A(+2) makes a break from G(+3) and stacks with G(+1). **c**, An  $\alpha$ -helix from CasC1 inserts into a cleft of CasE. **d**, The 5'-end repeat is buried by CasD, CasC6 and CasA. **e–g**, Expanded views of the interactions between the Cas proteins and the 5'-repeat segment. **h**, A loop of CasA is inserted into the cavity of CasD.





**Figure 3 | The spacer fragment is positioned in the continuous groove on the concave surface of CasC1–6.** **a**, The spacer fragment in red is not continuously stacked with five kinks present at regular intervals, namely at positions of the 6th, 12th, 18th, 24th and 30th nucleotide. Seven important  $\beta$ -hairpin arms from CasD, CasE and CasC2–6 subunits are highlighted. **b**, A  $\beta$ -hairpin of CasC3 inserts into the concave surface of CasC2. **c**, The interaction between the crRNA seed segment and CasC5–6. **d**, **e**, Structural comparison between CasC1 (cyan) and CasC2 (grey) (**d**), and between CasC5 (grey) and CasC6 (blue) (**e**).

(Extended Data Fig. 5d). Interestingly, because CasC6 is the last CasC subunit of the helical arrangement, it employs a CasD  $\beta$ -hairpin, even though CasD shares no similarity in overall folding with CasC.

Our structural analysis revealed an intriguing feature of a kink every 6 nucleotides within the crRNA spacer fragment. The 32-nucleotide spacer is embedded into a contiguous, positively charged groove formed by CasC1–6. The residues 1 to 5, 7 to 11, 13 to 17, 19 to 23 and 25 to 29 are continuously stacked. Intriguingly, these stacks are interrupted by five distinct breaks at steps 5–6, 11–12, 17–18, 23–24 and 29–30. The 6th, 12th, 18th, 24th and 30th residues flip out of the stacked helix, generating five regularly spaced kinks across the spacer, thus creating a discontinuous stacking arrangement of five similar fragments, each containing five stacked bases and one base splayed out (Fig. 3a, b). Interestingly, the kink positions precisely colocalize with those regions of the five  $\beta$ -hairpins projecting from CasC2–6 that cross the narrow gap at the break steps. Those two features are stabilized by the side-chains of F200<sup>CasC</sup> and V203<sup>CasC</sup> stacking with the respective two bases at each break site, and by hydrogen bonds formed between the  $\beta$ -hairpin and the crRNA phosphate backbone (Fig. 3b).

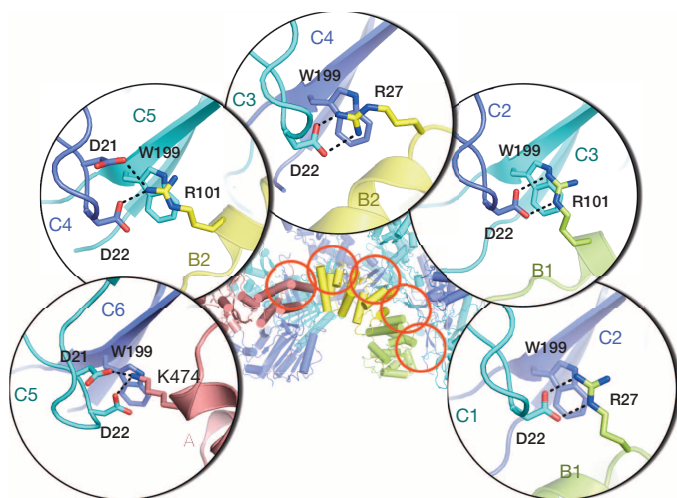
The 5-nucleotide stacked fragments between the kinks are positioned on the concave surface of the outer layer and are bound to their respective CasC motifs (Fig. 3a and Extended Data Fig. 2). Importantly, the 5-nucleotide segments are organized into an A-form helix, with the bases extending outwards, which is crucial for proper base pairing with target DNA. In contrast, the splayed out bases at the kink sites all face the distal domain of CasC, preventing any base pairing with target DNA, and most likely resulting in a non-continuous crRNA–DNA target duplex. Consistent with our structural finding, five site mismatches at the kink positions have a minimal effect on the target binding affinity, whereas mismatches positioned immediately upstream or downstream of the kink sites reduce the binding affinity by 20–30-fold (Extended Data Fig. 5e, f).

Residues 1–5 and 7–8, located within the first and second 5-nucleotide stacked segments of the spacer, are essential for target–DNA recognition<sup>20–22</sup>. In our structure, nucleotides 1–5, which are present on the surface of CasC6, are continuously stacked and extend outwards (Fig. 3c and Extended Data Fig. 6a). Residue 6 flips away from stack residues 1–5 by 180°, and therefore cannot pair with the DNA target, suggesting that it is not crucial for target recognition and explaining why this residue is not part of the seed sequence. Following this kink at position 6, the

base edges of residues 7–8 face in a similar direction as residues 1–5. Residues 7–8 continuously stack with residues 9–11, thus pairing between residues 7–8 and the target DNA should promote the formation of the crRNA–DNA target duplex, thereby facilitating the propagation to the 3' terminus of crRNA. Compared with the other 5-nucleotide stacked segments, partially obstructed by the CasB dimer, the first two segments containing the seed region are more accessible for the target strand. Moreover, the binding of proto-spacer adjacent motif (PAM)<sup>23,24</sup> to CasA allows for target recognition to start from the 5'-terminus of the spacer (Extended Data Fig. 6b–g). In conclusion, the fact that the seed bases face outwards towards solvent, together with the specific position near the tail end of the Cascade complex, allow for easy pairing with the DNA target, reminiscent of a short interfering RNA seed segment in Argonaute complexes<sup>25</sup>.

Structural comparison of the six CasC subunits revealed that, unlike the long  $\beta$ -hairpin arms of other CasC subunits, the corresponding region of CasC1 forms an  $\alpha$ -helix, rotates by  $\sim 90^\circ$  (Fig. 3d), and extends to the cleft of CasE (Fig. 2c and Extended Data Fig. 3d). F200<sup>CasC2–6</sup> and V203<sup>CasC2–6</sup> stack with the nucleotides at the break site, whereas F200<sup>CasC1</sup> and V203<sup>CasC1</sup> form hydrophobic interactions with CasE, thereby promoting Cascade complex formation. As observed in previous cryo-electron microscopy structures<sup>1,10</sup>, CasC6 adopts a different conformation from other CasC proteins. The rotation of the distal domain enables CasC6 to interact with CasD forming the 5' repeat segment binding pocket, and the particular conformation of the  $\beta 5$ – $\beta 6$  hairpin arm of CasD prevents CasC6 from taking on the same conformation as the other CasC subunits (Fig. 3e and Extended Data Fig. 5b).

The inner layer of Cascade is comprised of CasA and CasB1–2 subunits. As mentioned above, one loop of CasA inserts into the concave surface of CasD, while the C-terminal domain of CasA contacts CasB2 and CasC5–6 (Fig. 4 and Extended Data Fig. 7a, b). The elongated dimer of CasB1–2, located along the inner surface of the crRNA–CasC spine, is not involved in crRNA binding. Instead, it interacts with CasA, CasC and CasD, thus facilitating the Cascade assembly. Due to the curved shape of each CasB subunit, the asymmetric CasB–CasB interaction turns the CasB dimer into a curved structure as well, which is important for the interaction of CasB with three continuous CasC subunits (CasC1–3 with CasB1 and CasC3–5 with CasB2, Figs 1c, 4 and Extended Data Fig. 7c–f). Binding of the inner layer to the CasC helix is mediated by five unique binding spots, each of which is formed by a negatively charged



**Figure 4 | Assembly of CasA and CasB to the six-CasC helix mediated by the unique Asp-Arg/Lys-Trp triad.** The ribbon representation of the inner layer and the CasC helix is shown in the background. Each triad alignment is highlighted by red circle and the structural details are presented accordingly.

residue D22<sup>CasC</sup> that bridges to a positively charged residue (R27 or R101 of CasB1 or CasB2, as well as K474 of CasA), and is stabilized by stacking with the aromatic side chain of W199 of the following CasC molecule. Thus, the Asp-Arg/Lys-Trp triad unites two neighbouring CasC with either CasA or CasB (Fig. 4 and Extended Data Fig. 7).

Upon viral invasion, the Cascade complex binds double-stranded DNA exhibiting sequence complementary to the crRNA spacer sequence, and then recruits Cas3, a nuclease-helicase that cleaves the bound target DNA<sup>24,26–30</sup>. To understand the molecular basis for target recognition, we fitted each subunit structure into the cryo-electron microscopy map<sup>10</sup> and compared the structures with target bound and unbound. Overall, the outer layer displays negligible intrinsic motion, except that CasE rotates around the CasC1 binding site by  $\sim 15^\circ$  in an anti-clockwise direction, moving the crRNA stem-loop upward as a result (Extended Data Fig. 8a–d). In contrast, the inner layer undergoes a large conformational change upon target binding. While CasB1 and CasB2 move along the CasC helix for  $\sim 10$ – $12$  Å, CasA rotates by  $\sim 25^\circ$ , resulting in new interfaces of CasA–CasC and CasB–CasC (Extended Data Fig. 8c, e). Due to these conformational changes, the binding channel widens at the seed region, allowing the crRNA–target DNA duplex to fit in, and the motion of the CasB dimer stabilizing the target (Extended Data Fig. 8f, g).

In conclusion, our structural study has revealed that the crRNA plays an essential role not only in target recognition, but also in cascade complex assembly. Furthermore, the elongated  $\beta$ -hairpin loop from Cas protein represents a multifunctional element that serves a critical role in spacer RNA stabilization, CasC helix formation, and CasA–CasD as well as CasC–CasA and CasC–CasB interactions. Our structure also reveals a unique RNA-binding motif, in which a specific crRNA binding mode is used several times to assemble a multi-component complex.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 20 June; accepted 5 August 2014.

Published online 12 August 2014.

- Wiedenheft, B., Sternberg, S. H. & Doudna, J. A. RNA-guided genetic silencing systems in bacteria and archaea. *Nature* **482**, 331–338 (2012).
- Barrangou, R. *et al.* CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709–1712 (2007).
- Garneau, J. E. *et al.* The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* **468**, 67–71 (2010).
- Andersson, A. F. & Banfield, J. F. Virus population dynamics and acquired virus resistance in natural microbial communities. *Science* **320**, 1047–1050 (2008).

- Jansen, R., Embden, J. D., Gaastra, W. & Schouls, L. M. Identification of genes that are associated with DNA repeats in prokaryotes. *Mol. Microbiol.* **43**, 1565–1575 (2002).
- van der Oost, J., Jore, M. M., Westra, E. R., Lundgren, M. & Brouns, S. J. CRISPR-based adaptive and heritable immunity in prokaryotes. *Trends Biochem. Sci.* **34**, 401–407 (2009).
- Westra, E. R. *et al.* The CRISPRs, they are a-changin': how prokaryotes generate adaptive immunity. *Annu. Rev. Genet.* **46**, 311–339 (2012).
- van der Oost, J., Westra, E. R., Jackson, R. N. & Wiedenheft, B. Unravelling the structural and mechanistic basis of CRISPR-Cas systems. *Nature Rev. Microbiol.* **12**, 479–492 (2014).
- Makarova, K. S., Wolf, Y. I. & Koonin, E. V. The basic building blocks and evolution of CRISPR-cas systems. *Biochem. Soc. Trans.* **41**, 1392–1400 (2013).
- Wiedenheft, B. *et al.* Structures of the RNA-guided surveillance complex from a bacterial immune system. *Nature* **477**, 486–489 (2011).
- Jore, M. M. *et al.* Structural basis for CRISPR RNA-guided DNA recognition by Cascade. *Nature Struct. Mol. Biol.* **18**, 529–536 (2011).
- Reeks, J., Naismith, J. H. & White, M. F. CRISPR interference: a structural perspective. *Biochem. J.* **453**, 155–166 (2013).
- Brouns, S. J. *et al.* Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science* **321**, 960–964 (2008).
- Sashital, D. G., Jinek, M. & Doudna, J. A. An RNA-induced conformational change required for CRISPR RNA cleavage by the endoribonuclease Cse3. *Nature Struct. Mol. Biol.* **18**, 680–687 (2011).
- Gesner, E. M., Schellenberg, M. J., Garside, E. L., George, M. M. & Macmillan, A. M. Recognition and maturation of effector RNAs in a CRISPR interference pathway. *Nature Struct. Mol. Biol.* **18**, 688–692 (2011).
- Haurwitz, R. E., Jinek, M., Wiedenheft, B., Zhou, K. & Doudna, J. A. Sequence- and structure-specific RNA processing by a CRISPR endonuclease. *Science* **329**, 1355–1358 (2010).
- Nam, K. H. *et al.* Cas5d protein processes pre-crRNA and assembles into a cascade-like interference complex in subtype I-C/Dvul CRISPR-Cas system. *Structure* **20**, 1574–1584 (2012).
- Lintner, N. G. *et al.* Structural and functional characterization of an archaeal clustered regularly interspaced short palindromic repeat (CRISPR)-associated complex for antiviral defense (CASCAD). *J. Biol. Chem.* **286**, 21643–21656 (2011).
- Hrle, A. *et al.* Structure and RNA-binding properties of the type III-A CRISPR-associated protein Csm3. *RNA Biol.* **10**, 1670–1678 (2013).
- Wiedenheft, B. *et al.* RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions. *Proc. Natl Acad. Sci. USA* **108**, 10092–10097 (2011).
- Semenova, E. *et al.* Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. *Proc. Natl Acad. Sci. USA* **108**, 10098–10103 (2011).
- Künne, T., Swarts, D. C. & Brouns, S. J. Planting the seed: target recognition of short guide RNAs. *Trends Microbiol.* **22**, 74–83 (2014).
- Sashital, D. G., Wiedenheft, B. & Doudna, J. A. Mechanism of foreign DNA selection in a bacterial adaptive immune system. *Mol. Cell* **46**, 606–615 (2012).
- Hochstrasser, M. L. *et al.* CasA mediates Cas3-catalyzed target degradation during CRISPR RNA-guided interference. *Proc. Natl Acad. Sci. USA* **111**, 6618–6623 (2014).
- Wang, Y., Sheng, G., Juraneck, S., Tuschl, T. & Patel, D. J. Structure of the guide-strand-containing argonaute silencing complex. *Nature* **456**, 209–213 (2008).
- Westra, E. R. *et al.* CRISPR immunity relies on the consecutive binding and degradation of negatively supercoiled invader DNA by Cascade and Cas3. *Mol. Cell* **46**, 595–605 (2012).
- Mulepati, S. & Bailey, S. *In vitro* reconstitution of an *Escherichia coli* RNA-guided immune system reveals unidirectional, ATP-dependent degradation of DNA target. *J. Biol. Chem.* **288**, 22184–22192 (2013).
- Sinkunas, T. *et al.* *In vitro* reconstitution of Cascade-mediated CRISPR immunity in *Streptococcus thermophilus*. *EMBO J.* **32**, 385–394 (2013).
- Jackson, R. N., Lavin, M., Carter, J. & Wiedenheft, B. Fitting CRISPR-associated Cas3 into the helicase family tree. *Curr. Opin. Struct. Biol.* **24**, 106–114 (2014).
- Sinkunas, T. *et al.* Cas3 is a single-stranded DNA nuclease and ATP-dependent helicase in the CRISPR/Cas immune system. *EMBO J.* **30**, 1335–1342 (2011).

**Acknowledgements** We thank the staff at beamline BL-17U at Shanghai Synchrotron Radiation Facility (SSRF), and beamlines BL-1A, BL-5A and BL-17A at Photon Factory. The research was funded by Chinese Ministry of Science and Technology (2014CB910102 and 2011CBA01105), the Natural Science Foundation of China (31222014 and 31170705), and the Strategic Priority Research program of the Chinese Academy of Sciences (XDB08010203) to Y.W., and was supported by the Research Startup Fund from South University of Science and Technology of China and Shenzhen Government to Z.W. We thank D. Patel for assistance with manuscript editing, H. Wang, H. Li and F. Sun for experimental assistance, and T. Juelich for critical reading and editing of our manuscript.

**Author Contributions** H.Z., G.S. and M.W. expressed and purified and grew crystals of the Cascade complex. H.Z. and Y.W. collected x-ray diffraction data, J.W. made all constructs and did biochemical assays and Z.W., Y.W., G.B., H.Z. and W.G. solved the Cascade complex structure, Y.W. and Z.W. wrote the paper. All studies were undertaken under the supervision of Y.W.

**Author Information** Coordinates and structure factors have been deposited with Protein Data Bank accession code 4U7U. Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to Y.W. (ylwang@ibp.ac.cn) or Z.W. (wei.zy@sustc.edu.cn).



## METHODS

**Cloning, expression and purification of the Cascade complex.** The five Cas genes, *CasA*, *CasB*, *CasC*, *CasD* and *CasE*, were amplified from *E. coli* K12 genomic DNA by polymerase chain reaction (PCR) and subcloned into different expression vectors. *CasA* was inserted into pET22b (Novagen, Amp<sup>R</sup>). *CasB* was inserted into pRSFDuet-1 (Novagen, Kan<sup>R</sup>) containing an N-terminal His<sub>6</sub>-tag. *CasC*, *CasD* and *CasE* were inserted into a modified pCDFDuet-1 (Novagen, Str<sup>R</sup>). Template CRISPR was chemically synthesized from Sangon Biotech (sequences provided below) and subcloned into pACYCDuet-1 (Novagen, Chlor<sup>R</sup>). Generation of the CasD and CasC point mutants was performed by site-directed mutagenesis. All constructs were confirmed by sequencing. All four vectors were then co-transformed into the *E. coli* BL21 (DE3). Overexpression of Cascade was induced with 0.2 mM isopropyl-β-D-thiogalactoside (IPTG) at an OD<sub>600 nm</sub> of 0.4. After growing for 20 h at 16 °C, cells were harvested by centrifugation, homogenized in 1× PBS buffer A (pH 7.3) containing 5% glycerol. After sonication and centrifugation, supernatants were incubated with Ni Sepharose resin (GE Healthcare) for 2 h and eluted with an increasing imidazole gradient in buffer A. Then, the recombinant complex was further purified by Heparin HP column (GE Healthcare) and MonoQ 5/50 (GE Healthcare) with buffer B (20 mM HEPES, pH 7.5, 100 mM NaCl) and buffer C (20 mM HEPES, pH 7.5, 1 M NaCl). For crystallization trials, the protein complex was purified by gel filtration chromatography (Superdex 200 10/300 GL, GE Healthcare) in buffer D (20 mM HEPES, pH 7.5, 10 mM DTT, 1 mM EDTA). The peak fractions were collected and concentrated to about 4 mg ml<sup>-1</sup> for crystallization. The entire purification procedure was performed at 4 °C and purity of the Cascade complex was verified by SDS-PAGE.

**Crystallization and data collection of Cascade complex.** Crystals of the Cascade complex were grown at 16 °C using the hanging drop vapour diffusion method with a reservoir solution containing 100 mM Tris-HCl, pH 8.5, 200 mM NaAc, 100 mM Glycine, 50 mM LiAc and 11–13% (w/v) polyethylene glycol (PEG) 4,000. Crystals appeared after 4 days and grew to full size within 2 weeks. For data collection, crystals were transferred into mother liquor supplemented with 20% (v/v) glycerol and flash-cooled in liquid nitrogen. Diffraction data were measured and collected at 100 K with wavelength of 1.1 Å at beamline BL1A, BL5A and BL17A of Photon Factory (Japan, KEK) or at beamline BL17U of Shanghai Synchrotron Radiation Facility (SSRF). All data sets were processed using HKL2000 or iMosflm<sup>31,32</sup>. The crystals were identified as belonging to space group P1, with two Cascade complexes per asymmetric unit.

**Structure determination and refinement of the Cascade complex.** By using the 8.8-Å electron microscopy map of the target-unbound Cascade complex as a search model, we found an optimal solution for molecular replacement in PHASER<sup>33,34</sup>. To improve the phase the non-crystallographic symmetries (NCS) for two Cascade complexes and for five CasC molecules were identified and applied in density modification process using RESOLVE<sup>35</sup>. The phase was dramatically improved with the powerful NCS average and then was gradually extended to 3.5 Å (Extended Data Fig. 9). The structural model was manually built and refined in Phenix<sup>36</sup> against the 3.05 Å data set. COOT<sup>37</sup> was used for model building and adjusting. The final model was validated using MolProbity<sup>38</sup> and Procheck<sup>39</sup>. The refinement statistics are listed in Extended Data Table 1. All structure figures were prepared using PyMOL (<http://www.pymol.org/>). Individual Cas proteins are colour-coded. The repeat segments of crRNA are in orange and the spacer is in red.

**Cascade mutagenesis and gel filtration assay.** Two types of Cascade mutants, CasCΔ and CasDΔ, were produced. In CasCΔ, residues Phe 200, Asp 204, Asp 205 and Leu 206 of CasC were mutated to Ala. In CasDΔ, residues 75–104 of CasD were replaced by a (GGG)<sub>4</sub> linker. CasCΔ and CasDΔ were cloned and expressed using the same strategy as used for wild-type Cascade. Both mutants were purified using Ni-affinity resin and analysed by gel filtration assay (Superdex200 10/300 GL, GE Healthcare). The assays were performed in a buffer containing 20 mM HEPES pH 7.5 and 100 mM NaCl. The relevant fractions were then analysed by SDS-PAGE and visualized by Coomassie blue staining.

**pACYC-target invasion assay.** For our invasion assay, we used a CRISPR system containing a 32-nucleotide spacer: 5'-GGCTCCCTGTCGGTTGTAATTGATAA TGTTGA-3'. The complementary sequence of the spacer (5'-TCAACATTATC AATTACAACCGACAGGGAGCC-3') followed by a PAM sequence (ATG) was constructed into pACYC-1 between the restriction sites NdeI and XhoI, yielding plasmids that contain an artificial protospacer target (pACYC-target). *E. coli* BL21

(DE3) cells that contain CRISPR-encoding plasmids with either the wild-type or mutant Cascade genes together with Cas3 were inoculated in LB containing kanamycin (50 μg ml<sup>-1</sup>), ampicillin (100 μg ml<sup>-1</sup>) and streptomycin (50 μg ml<sup>-1</sup>), and grown to an OD<sub>600 nm</sub> of 0.3. Expression of both Cas genes and CRISPR was induced for 45 min with 0.5 mM IPTG. Cells were then collected at 4 °C and made competent according to standard protocols. Transformation was performed by adding either 60 ng pACYC-target or 60 ng pACYC.

Next, cells were grown in LB for 60 min at 37 °C before plating on LB-agar plates containing 0.5 mM IPTG, ampicillin (100 μg ml<sup>-1</sup>), kanamycin (50 μg ml<sup>-1</sup>), streptomycin (50 μg ml<sup>-1</sup>) and chloramphenicol (34 μg ml<sup>-1</sup>). Plates were incubated for 12 h at 37 °C before observation.

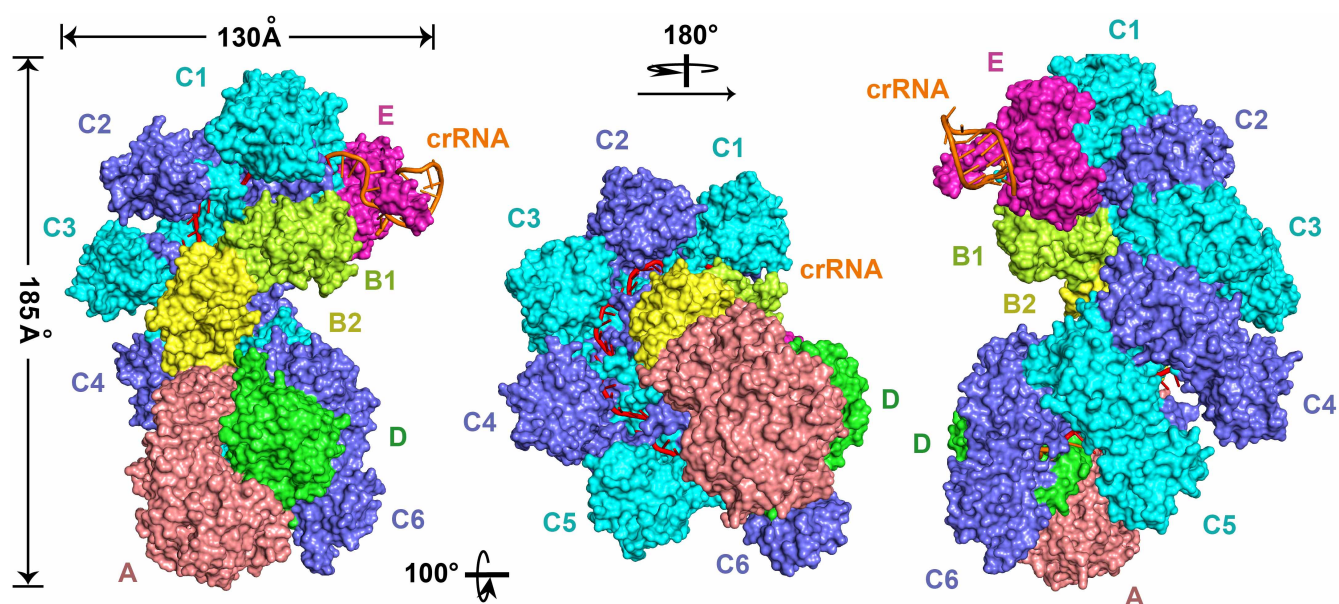
The pACYC control without target sequence did not trigger CRISPR/Cas interference, and was therefore retained, with no inhibition of growth observed. In contrast, the plasmid pACYC-target containing target sequence was recognized and degraded by CRISPR/Cas system, growth inhibition observed on selection medium.

**Plasmid loss assay.** *E. coli* BL21 (DE3) strains expressing either wild-type Cascade (WT) or Cascade complex in which the hairpin-arm of CasD was replaced with a (GGG)<sub>4</sub> linker (DΔ), were transformed with pACYC-target plasmid. Cultures were inoculated in LB containing kanamycin (50 μg ml<sup>-1</sup>), ampicillin (100 μg ml<sup>-1</sup>), streptomycin (50 μg ml<sup>-1</sup>) and chloramphenicol (34 μg ml<sup>-1</sup>), and grown to an OD<sub>600 nm</sub> of 0.6. Expression of cas genes and CRISPR was induced for 5 h with 0.5 mM IPTG. Cells were serially diluted and plated on LB-agar non-selective plates (containing 50 μg ml<sup>-1</sup> kanamycin, 100 μg ml<sup>-1</sup> ampicillin and 50 μg ml<sup>-1</sup> streptomycin), or selective plates containing additional 34 μg ml<sup>-1</sup> chloramphenicol. All plates were incubated for 12 h at 37 °C before observation. If the plasmid can trigger CRISPR/Cas interference, it will be recognized and degraded. As a consequence, growth will be inhibited on selection medium. Alternatively, the plasmid will remain intact, and colonies will appear on selection medium.

**Isothermal titration calorimetry (ITC).** DNA oligonucleotides for ITC were dissolved in buffer containing 20 mM HEPES, pH 7.5, and 100 mM NaCl. The Cascade complex for ITC measurements was purified by Superdex 200 with HEPES buffer (20 mM, pH 7.5) containing 100 mM NaCl. The purified Cascade complex (1 μM) was loaded into the cell, and the ssDNA solution (10 μM) in the syringe. Experiments were carried out at 20 °C using a MicroCal ITC200 calorimeter (GE Healthcare). The titration data were processed and fitted using Origin 7.0 software.

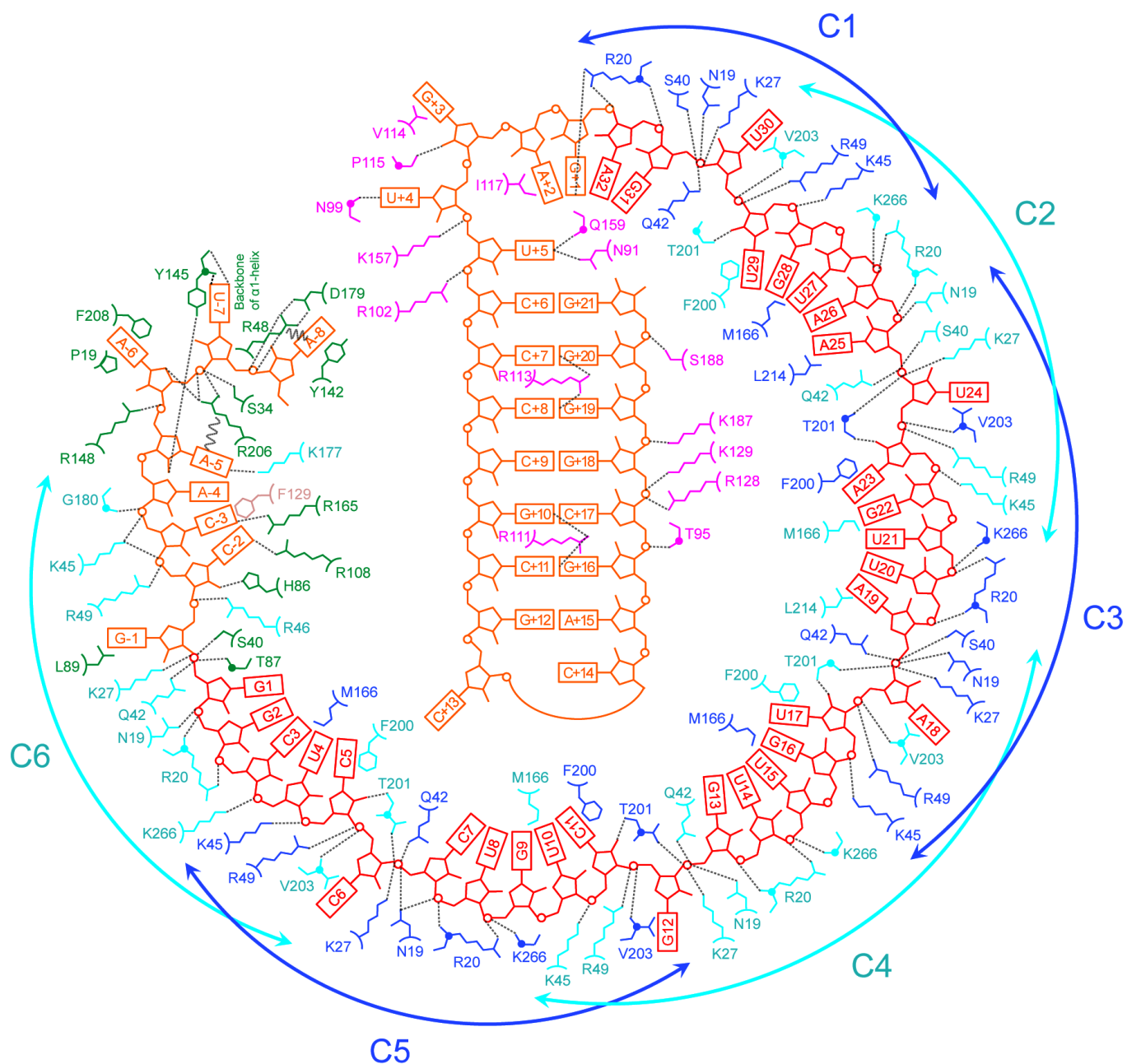
**Template CRISPR sequence (from 5' to 3').** GAGTCCCCGCGCCAGCGGGG ATAAACCGGGCTCCCTGTCGGTTGTAATTGATAATGTTGAGAGTTCCC CGCGCCAGCGGGGATAAACCGGGCTCCCTGTCGGTTGTAATTGATAAT GTTGAGAGTTCCCCGCGCCAGCGGGGATAAACCGGGCTCCCTGTCGG TTGTAATTGATAATGTTGAGAGTTCCCCGCGCCAGCGGGGATAAACCG GGCTCCCTGTCGGTTGTAATTGATAATGTTGAGAGTTCCCCGCGCCAG CGGGGATAAACCGGGCTCCCTGTCGGTTGTAATTGATAATGTTGAGAG TTCCCCGCGCCAGCGGGGATAAACCGGGCTCCCTGTCGGTTGTAATTG ATAATGTTGAGAGTTCCCCGCGCCAGCGGGGATAAACCGGGCTCCCT GTCGGTTGTAATTGATAATGTTGAGAGTTCCCCGCGCCAGCGGGGATA AACCG.

- Battye, T. G., Kontogiannis, L., Johnson, O., Powell, H. R. & Leslie, A. G. iMOSFLM: a new graphical interface for diffraction-image processing with MOSFLM. *Acta Crystallogr. D* **67**, 271–281 (2011).
- Otwinski, Z. & Minor, W. Processing of X-Ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326 (1997).
- Xiong, Y. From electron microscopy to X-ray crystallography: molecular-replacement case studies. *Acta Crystallogr. D* **64**, 76–82 (2008).
- McCoy, A. J. et al. Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).
- Terwilliger, T. C. Maximum-likelihood density modification. *Acta Crystallogr. D* **56**, 965–972 (2000).
- Adams, P. D. et al. PHENIX: building new software for automated crystallographic structure determination. *Acta Crystallogr. D* **58**, 1948–1954 (2002).
- Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60**, 2126–2132 (2004).
- Chen, V. B. et al. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66**, 12–21 (2010).
- Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **26**, 283–291 (1993).

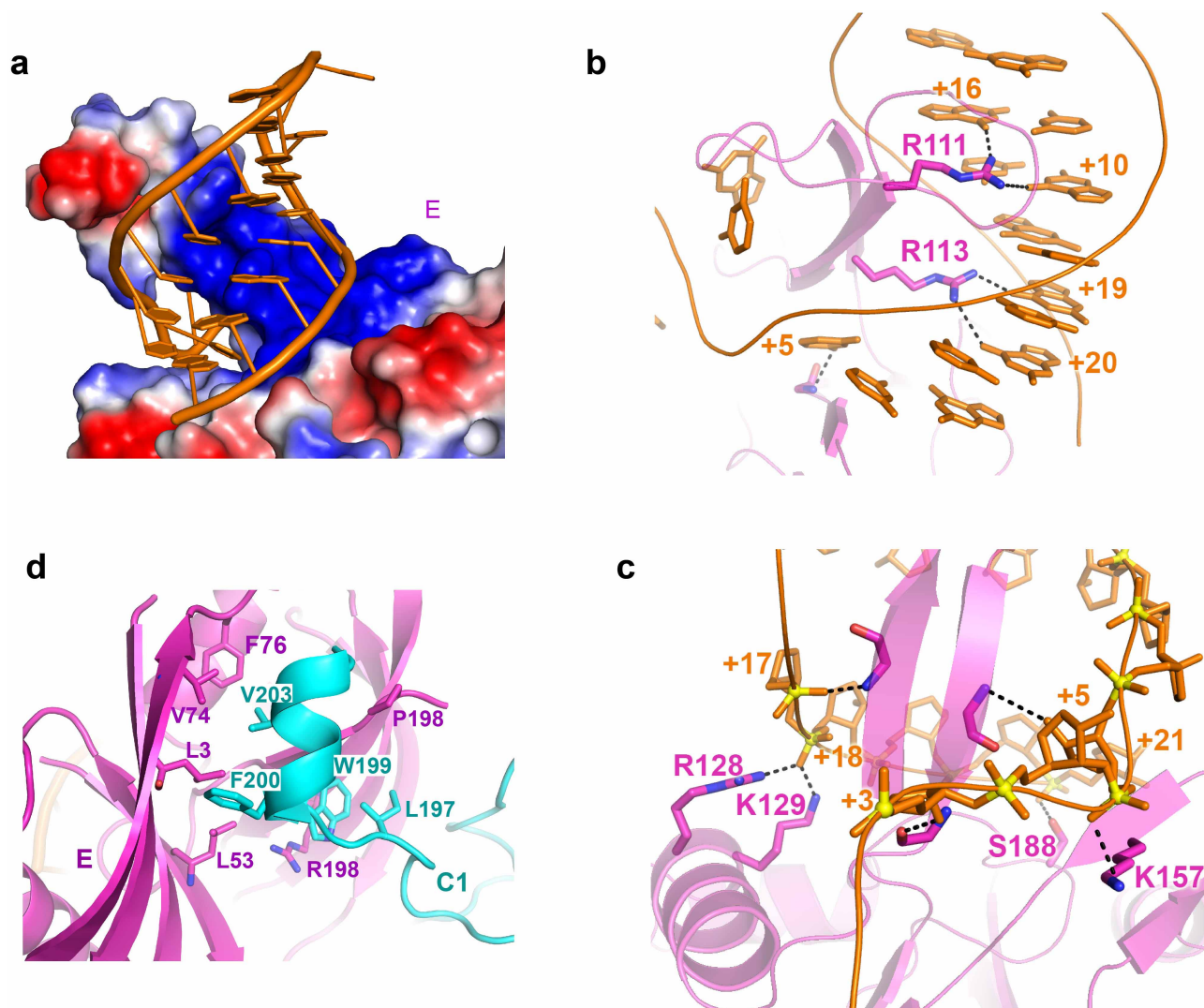


Extended Data Figure 1 | Overall surface view of the Cascade complex with the same orientations as those used for Fig. 1b.





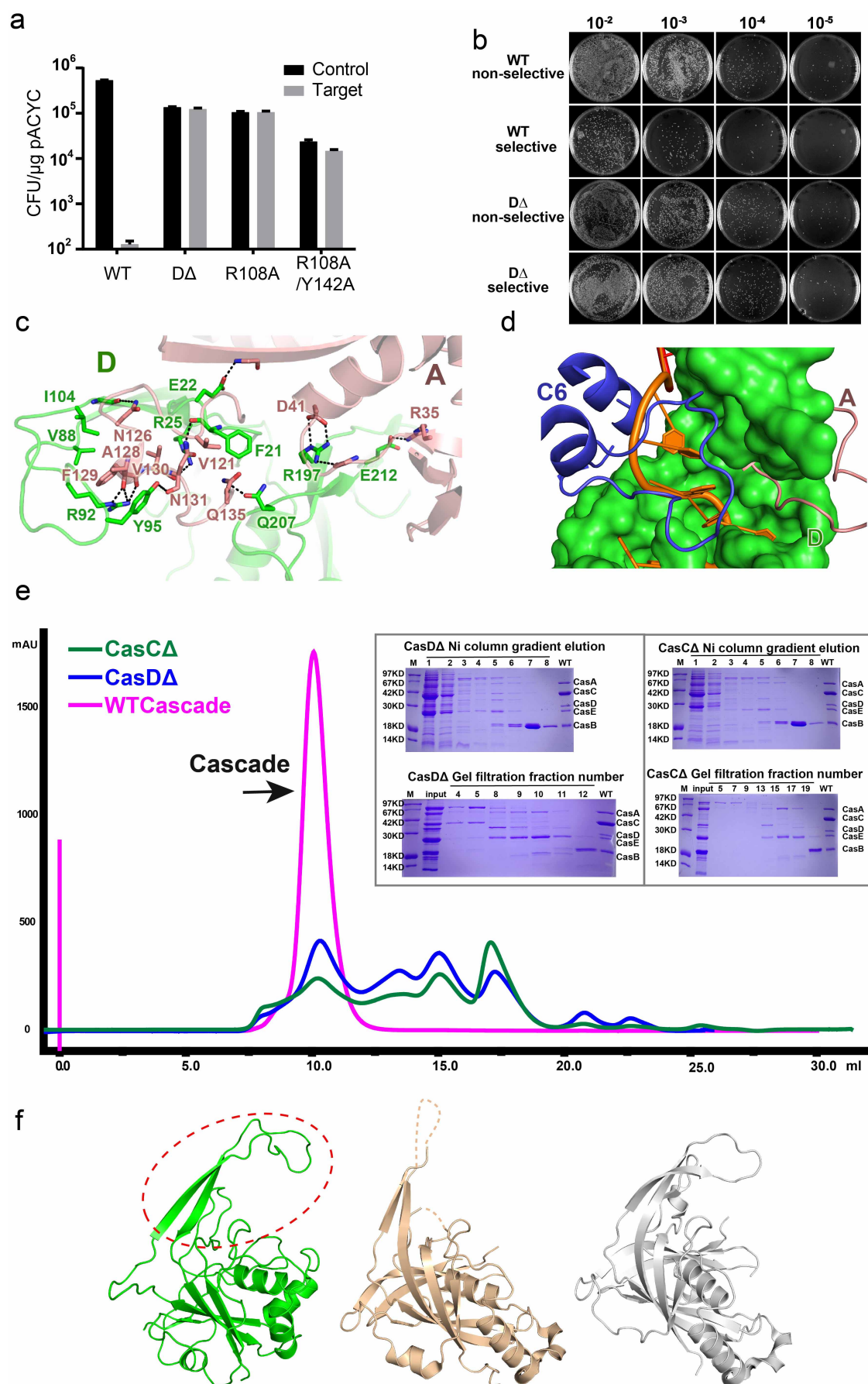
**Extended Data Figure 2 | Schematic summary of Cas-protein-crRNA interactions.** Hydrogen bonds and salt bridges are indicated by dashed lines. Cation- $\pi$  interactions are indicated by wavy lines.



**Extended Data Figure 3 | Structure of CasE subunit in the Cascade complex.**

**a**, CasE is shown as a surface representation, and is labelled according to electrostatic potential (red, negative charge; blue, positive charge), and RNA is shown in ribbon representation (orange). **b**, **c**, Magnified view of the

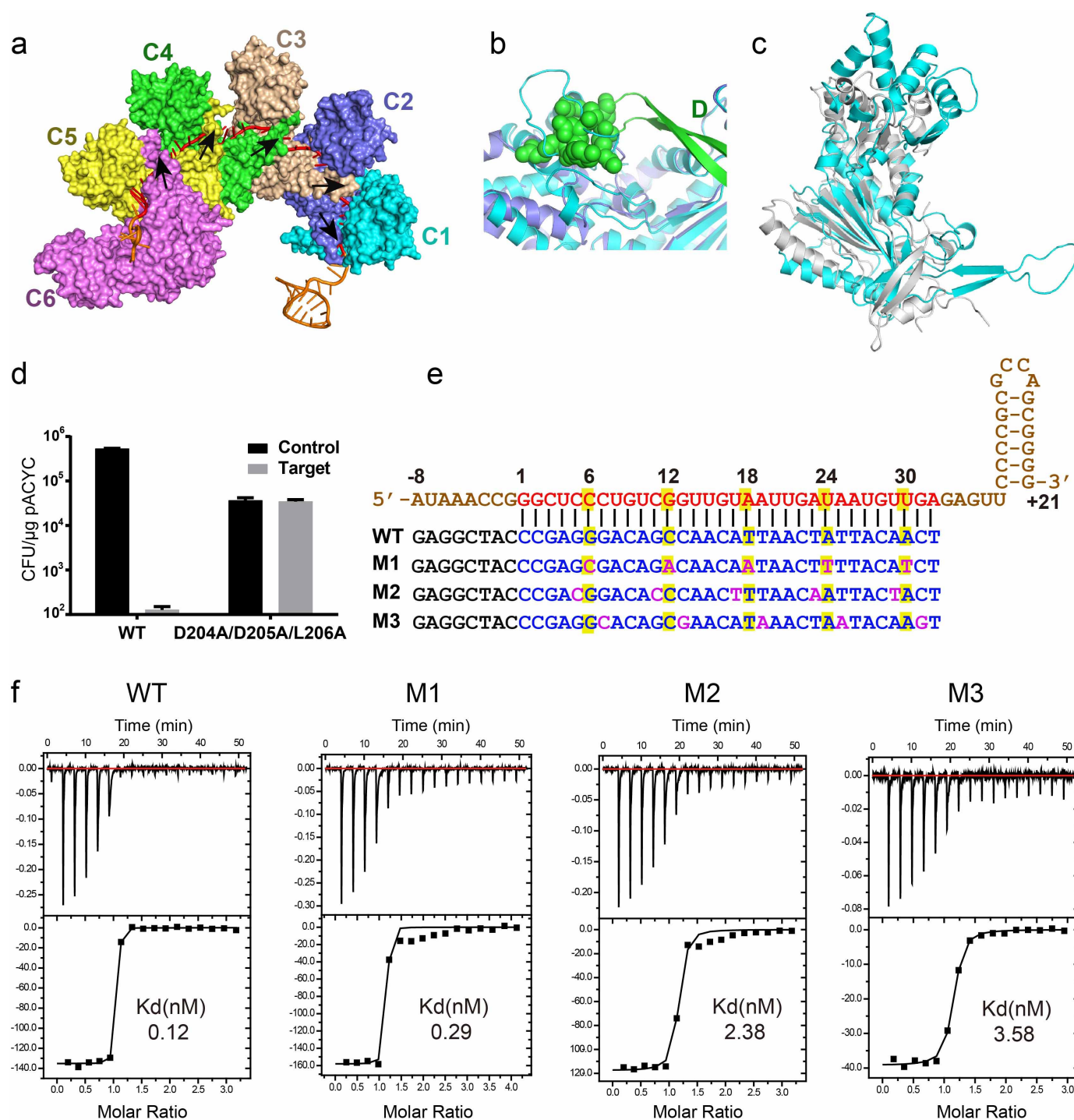
sequence-specific interaction between CasE and the major groove of 3'-end repeat (**b**), and the interaction between CasE and 3'-end repeat backbone (**c**). **d**, Expanded view of the interaction between CasE and CasC1.



**Extended Data Figure 4 | The  $\beta$ -hairpin arm of CasD is critical for Cascade function.** **a**, pACYC-target invasion assay. Competent *E. coli* BL21 (DE3) cells expressing either wild type Cascade (WT) or a mutant form in which the hairpin-arm was replaced with a (GGG)<sub>4</sub> linker CasD mutant (D $\Delta$ ), or R108A, R108A/Y142A CasD mutants, were transformed with either pACYC plasmid (Control) or pACYC-target plasmids (Target). Colony-forming units per microgram pACYC (CFU per  $\mu$ g) are depicted for each of the strains. The statistics was based on 10 replicates. Error bars represent the standard error of the mean (s.e.m.). Upon transformation with pACYC plasmid control, *E. coli* cells expressing WT Cascade exhibited much higher transformation efficiencies than cells transformed with pACYC-target plasmid. However, the transformation efficiencies were high in *E. coli* cells expressing mutant CasD and transformed with either pACYC control or pACYC-target. These results show that the hairpin-arm of CasD is essential for proper function of the Cascade complex. **b**, Plasmid loss assay. *E. coli* BL21 (DE3) cells expressing wild-type or casD $\Delta$ (75–104)-replaced Cascade complex and pACYC-target, were serially diluted and grown on either non-selective or selective plates upon induction with 0.5 mM IPTG. Growth of BL21 cells expressing WT Cascade was seriously inhibited on selection medium compared with growth on non-selection medium. However, BL21 cells with mutant CasD displayed similar

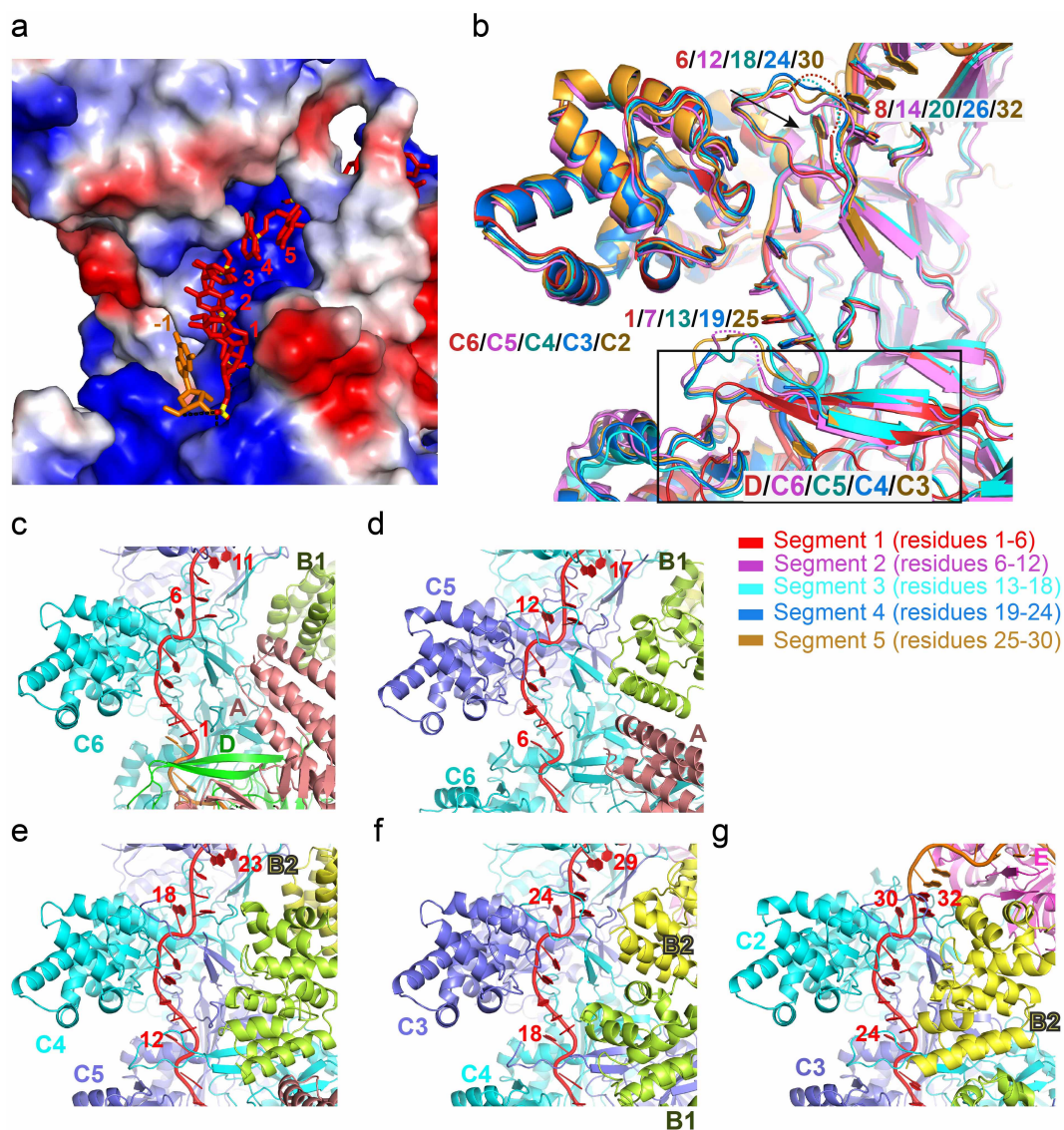
growth on selection and non-selection medium, indicating that the target sequence was retained upon introduction of the CasD mutant. **c**, Interface between CasD (green) and CasA (salmon). **d**, Interface between CasD (green) and CasC6 (blue). **e**, Pull-down and gel-filtration analysis of the Cascade assembly. WT Cascade complex was formed effectively in the Ni-NTA column and was eluted out with elution buffer containing 100 mM imidazole. In contrast, very limited amounts of the co-expressed Cascade complex with the D $\Delta$  mutant or C $\Delta$  mutant were obtained in the 100 mM imidazole elution, with the majority of the Cas proteins were in the flow-through. The fractions eluted with 100 mM imidazole were collected and applied for the analytical gel filtration assay. Compared with the WT Cascade, both D $\Delta$  mutant and C $\Delta$  mutant Cascade did not form a stable complex in gel filtration, further confirming the critical role of the  $\beta$ -hairpin arm of CasD and CasC for the assembly of the Cascade complex. The CasC mutant (C $\Delta$ ) contains F200A mutation as well as replacement mutation, where residues D204 to L206 were replaced by Ala. **f**, Structural comparison among *E. coli* CasD (green) and Cas5d in the RNA-free state of *Streptococcus pyogenes* (wheat, PDB id: 3VZH) and *Bacillus halodurans* (grey, PDB id: 4F3M). The  $\beta$ -hairpin arm of CasD is highlighted by a red dashed circle.





**Extended Data Figure 5 | The unique multi-kinked conformation within the crRNA spacer region.** **a**, The six CasC subunits form a right-handed helix, with a groove for crRNA binding. Except for CasC1, the β-hairpin (highlighted by a black arrow) of each CasC molecule extends up into the groove of its preceding CasC, and thus results in five kinks observed for the crRNA spacer. **b**, Superposition of CasC5 (cyan) and CasC6 (blue). The top of long β-hairpin in CasD is shown as green spheres, depicting the clash between the distal domain of CasC6 and the long β-hairpin top given it adopts the same conformation as CasC5. **c**, CasC (cyan) shares a similar overall fold with *Sulfolobus solfataricus* Cas7 (grey, PDB id 3PS0). **d**, The pACYC-target invasion assay showing that mutating three interacting residues (D204<sup>CasC</sup>,

D205<sup>CasC</sup>, L206<sup>CasC</sup>) at the β-hairpin arm largely impairs Cascade activity. **e**, Four designed DNA targets with complementary sequence (WT) to the spacer segment of crRNA or with non-complementary mutations at kinked sites (M1), at immediately upstream nucleotides (M2), and at immediately downstream nucleotides (M3). The five kinked sites are highlighted as yellow background, with the mutated sequences in purple. **f**, ITC-based analysis of the Cascade complex and four DNA targets interactions. ITC analysis showed that the mutations of the flipping of these residues had no obvious effect on target recognition. However, mutations in either the first or the last nucleotides in the 5-nucleotide stacked region largely affected target binding.

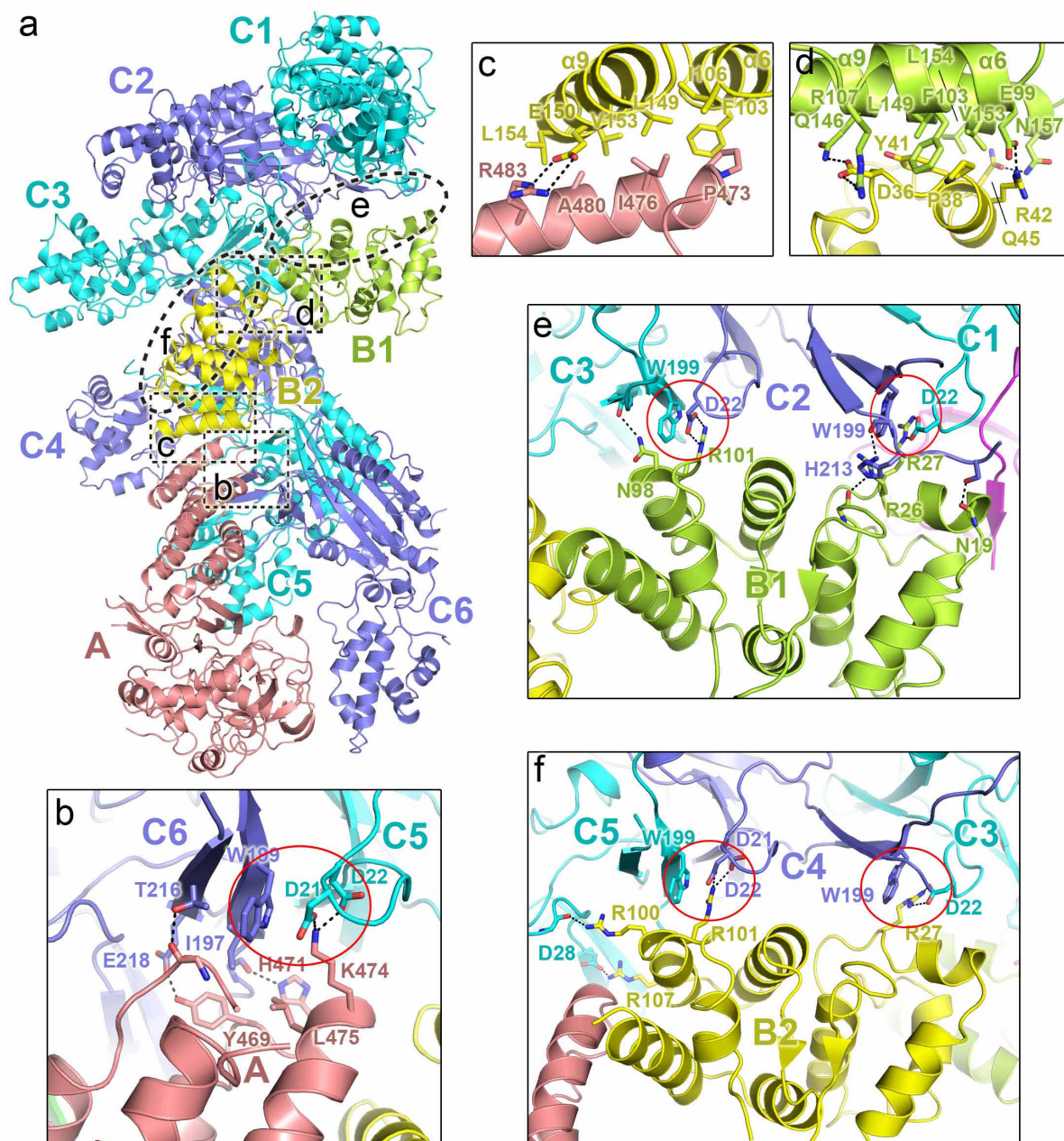


#### Extended Data Figure 6 | Structural analysis of the crRNA spacer segments.

**a**, The first 5-nucleotide segment in the spacer region is positioned in the positively charged groove of CasC. **b**, Overlap of the five 5-nucleotide segments within the crRNA spacer region and their interacting Cas proteins. The 5-nucleotide segments are interrupted by single flipped-out bases, which are indicated by a black arrow, thereby resulting in kinks at these positions. The conformations of the five segments together with the kink sites are essentially

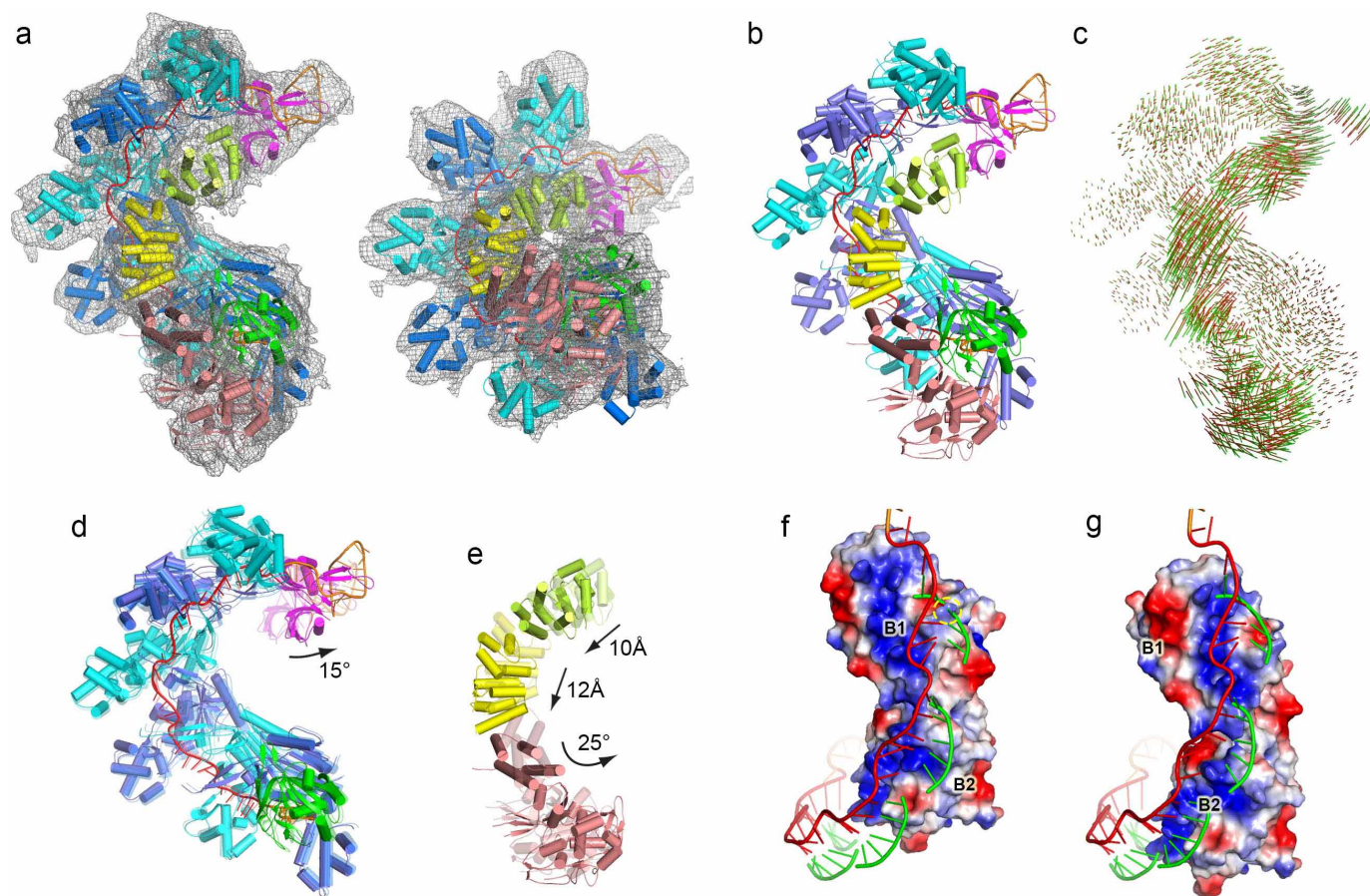
identical. Like those of CasC2–6, the  $\beta$ -hairpin arm of CasD is involved in maintaining the kink at G(–1). Interestingly, although the overall foldings are distinct between CasD and CasC, their  $\beta$ -hairpin arms share similar conformations, which are highlighted by a black box. **c–g**, Structural comparison of the five 5-nucleotide stacked segments with Cas proteins shown in the same orientation. The Cascade complex is colour-coded by chains as shown in Fig. 1.





**Extended Data Figure 7 | Inter-subunit interactions required for Cascade assembly.** **a**, Ribbon representation of the organization of CasA and CasB dimers and their associations with the CasC helix. The inter-subunit interfaces are highlighted with dashed-line boxes. **b–f**, The structural details of the

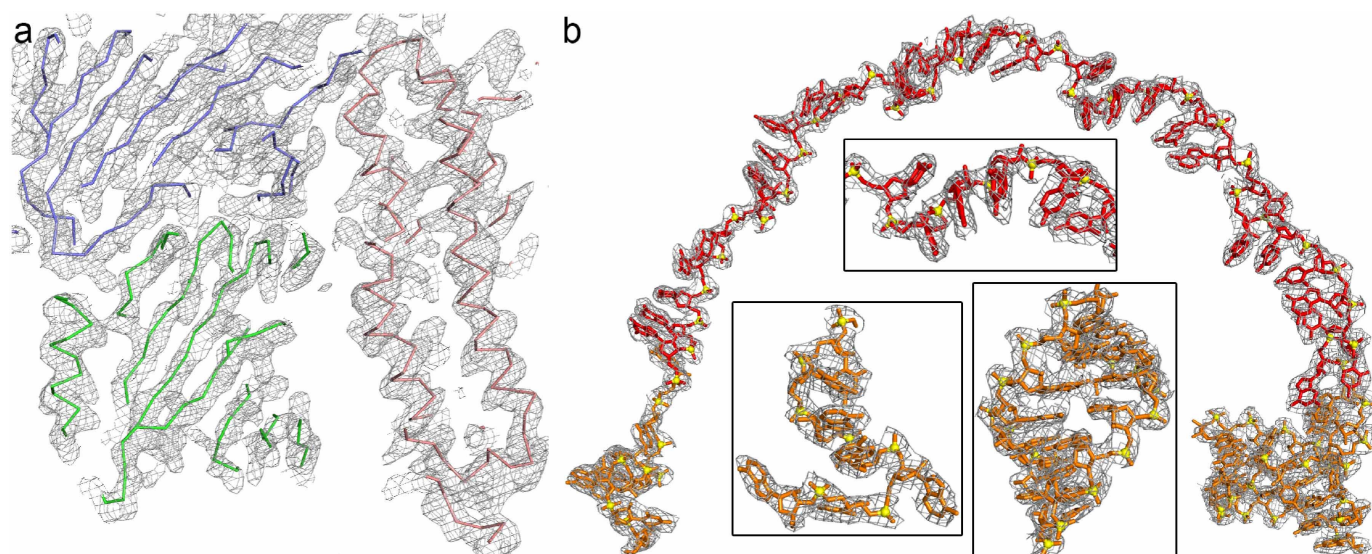
CasA–CasC (**b**), CasA–CasB2 (**c**), CasB1–CasB2 (**d**), CasB1–CasC (**e**) and CasB2–CasC (**f**) interactions. The five triads, which connect CasA and the CasB dimer to the CasC helix, are highlighted with red circles. Hydrogen bonds and salt bridges are indicated by dashed lines.



**Extended Data Figure 8 | Structural comparison between target-bound and unbound Cascade.** **a**, The building of the structural model of the target-bound Cascade. The subunits were individually docked into the 9Å-electron microscopy map. The fitted model is shown in two views with the electron microscopy density overlapping (contoured at  $2\sigma$ ). **b**, Ribbon representation of the target-bound Cascade. **c**, Subunit motions. The movement was represented by red-green lines, which were drawn by connecting each C $\alpha$  atom (green) in the target-bound Cascade to corresponding C $\alpha$  atom (red) in the unbound Cascade after the two Cascade structures were superimposed. Thus, the lengths of the lines correlate with the motion scale. The motions of the outer and inner layers are shown by superimposing the target-bound with unbound Cascade structure in **d** and **e**, respectively. The target-free structures

are rendered semi-transparent. The translational and rotational motions are indicated by arrows showing the motional direction with translation distances and rotation angles labelled, respectively. **f**, **g**, Comparison of electrostatic surface potentials for the CasB dimer in the target-free (**f**) and target-bound Cascade (**g**) with cartoon representations of the crRNA spacer and its paring target (green). A highly positively charged groove of the CasB dimer fits well with the negatively charged target backbone in **g**, but not in **f**. Also, the paring target partially clashes with the CasB dimer in **f**, with the clashing site highlighted by a yellow circle. The structural analysis suggests that the motion of the CasB dimer identified in the target-bound Cascade complex is required for proper target binding.





**Extended Data Figure 9 | Electron density maps showing the model quality.**

**a**, The 3.5 Å-electron density map contoured at  $1.5\sigma$  generated from the phase after density modification treatment with the final structural model superimposed. C $\alpha$  backbones are shown in with the same colour-coding as used

in Fig. 1. **b**, The  $2F_o - F_c$  map contoured at  $1\sigma$  of the whole crRNA is shown with the refined RNA model superimposed. Zoom-in views of the 5'-end, 3'-end and the 5-nucleotide stacked segment are shown in the lower left, lower right and upper insets, respectively.

Extended Data Table 1 | Statistics of data collection and refinement

	Cascade
Data collection	
Space group	P1
Cell dimensions	
a, b, c (Å)	111.41, 118.14, 225.87
$\alpha$ , $\beta$ , $\gamma$ (°)	92.24, 93.55, 106.06
Wavelength (Å)	1.1
Resolution (Å)	50-3.05 (3.16-3.05)
R <sub>merge</sub> (%)	9.1 (40.5)
I/ $\sigma$	11.7 (2.1)
Completeness (%)	98.8 (98.3)
Redundancy	3.05 (2.9)
Refinement	
Resolution (Å)	45-3.05
No. reflections	195,250
R <sub>work</sub> / R <sub>free</sub> (%)	16.4/20.7
No. atoms	
Protein	53257
RNA	2596
Average B (Å <sup>2</sup> )	
Protein	43.4
RNA	45.6
R.m.s. deviations	
Bonds (Å)	0.006
Angle (°)	1.2
Ramachandran statistics (%)	
Most favorable	89.8
Additionally allowed	9.8
Generously allowed	0.3
Disallowed	0.1

Values in parentheses are for the highest resolution shell.

# INTERACTIVE NOTEBOOKS: SHARING THE CODE

*The free IPython notebook makes data analysis easier to record, understand and reproduce.*

ILLUSTRATION BY THE PROJECT TWINS



BY HELEN SHEN

Flying high above the Pacific Ocean, Titus Brown is taking a deep dive into his students' research code. The long journey from Michigan State University in East Lansing to a conference in Melbourne, Australia, provides the perfect chance for the bioinformatician to scrutinize his lab's new algorithm for removing errors from RNA sequencing data.

Three years ago, Brown might have waited until he was back in his office. It is difficult to dig into other researchers' code without them being present to explain it, make changes and produce updated results. But these days, Brown can work with his lab from afar using a free, open-source software package called IPython, which helps

researchers to keep a detailed lab notebook for their computational work.

Brown's students write explanatory text and intersperse it with raw code and the charts and figures that their algorithms generate. Sitting in the aeroplane with an IPython notebook downloaded to his laptop, Brown can interact with the work. He tweaks and re-runs the code, which executes directly in the document he is reading — allowing him to see instantly whether his changes are improving the algorithm. "I can go through their notebook, understand exactly what they did and modify it, explore different parameters and look at different views," he says. "I can do this from anywhere in the world."

Designed to make data analysis easier to share and reproduce, the IPython notebook is being

used increasingly by scientists who want to keep detailed records of their work, devise teaching modules and collaborate with others. Some researchers are even publishing the notebooks to back up their research papers — and Brown, among others, is pushing to use the program as a new form of interactive science publishing.

## BETTER BOOKKEEPING

The IPython notebook was developed in 2011 by a team of researchers led by Fernando Pérez, a data scientist at the University of California, Berkeley, and computational physicist Brian Granger at California Polytechnic State University in San Luis Obispo. "We built it by solving problems that we ourselves had as researchers and educators," says Pérez. ►

## PROGRAMMING

*IPython for beginners***Getting the tools**

An interactive *Nature* demonstration of an IPython notebook is available at [go.nature.com/ohkjks](http://go.nature.com/ohkjks). The software for the notebook can be downloaded from the IPython website at [go.nature.com/mq8nip](http://go.nature.com/mq8nip).

**Learning the ropes**

Instructions for proficient coders can be found at [go.nature.com/sdbolb](http://go.nature.com/sdbolb); example notebooks are at [go.nature.com/awtkxn](http://go.nature.com/awtkxn).

For people unfamiliar with Python, a host of resources exist online. OpenTechSchool's educational unit on data analysis with Python ([go.nature.com/gpuypx](http://go.nature.com/gpuypx)) includes an introduction to the IPython notebook.

For hands-on training, Software Carpentry, a volunteer organization that teaches basic software skills, can help. The group will run two-day workshops on the IPython notebook by invitation throughout the world ([go.nature.com/fj6sza](http://go.nature.com/fj6sza)).

**Sharing the goods**

IPython notebooks can be shared in online repositories such as GitHub, or over e-mail. Recipients need IPython software to view and edit the notebooks. Notebook authors can also use a program called nbviewer to create an online version of their notebook that will be viewable, but not editable ([go.nature.com/ry6g4j](http://go.nature.com/ry6g4j)).

► Pérez and Granger saw that data scientists found it hard to share detailed but understandable descriptions of their raw code that would allow others to build on their research. That is partly because many scientists in computation-intensive fields write code in an iterative and piecemeal fashion as each analysis reveals new insight and spins off multiple lines of inquiry. Keeping track of the different versions of code that produce various figures, and linking those files with explanatory notes, is a headache. And what gets published is usually not detailed enough for the reader to follow up on. “In my own computational physics work,” says Granger, “a high-level description of the algorithm that goes into the paper is light years away from the details that are written in the code. Without those details, there is no way that someone could reproduce it in a reasonable time scale.”

The IPython notebook addresses both issues by helping scientists to keep track of their work, and by making it easy to share and for others to explore the code. The ‘I’ in IPython refers to an ‘interactive’ command window that helps users to run code, access variables, call up data analysis packages and view plots, while the Python refers to the popular programming language that the notebook is based on. (Pérez, Granger and their colleagues are now moving the notebook into a project called Jupyter, which aims to make IPython more compatible with other languages, including Julia and R).

**CODE CHOPS**

At the University of Texas at Austin, Tal Yarkoni uses the IPython notebook to run automated meta-analyses of brain imaging studies to uncover patterns of neural activity involved in language processing, emotion and other processes. The psychoinformatician plans to publish the notebooks as companions to his future journal articles. “The more complicated the analyses, the greater the benefits of being able to

convey all that in one simple document,” he says.

Applications similar to the IPython notebook already exist for various programming languages. Mathematica and Maple — commercial analysis packages popular among mathematicians — include notebooks or notebook-like programs. MATLAB, a commercial package used heavily in signal processing, engineering and medical-imaging research, also supports a notebook application. Each of these notebooks is specialized for its corresponding proprietary programming language.

A number of notebooks and notebook-like programs exist in the open-source world; knitr works with the R coding language, which is especially powerful for statistical analysis. And the Sage mathematical software system, which is also based on the Python language, supports its own notebook. DEXY is a notebook-like program that focuses on helping users to generate papers and presentations that incorporate prose, code, figures and other media.

But the IPython notebook has become one of the most widely adopted programs of its kind, says Ana Nelson, the creator of DEXY. “So many people have heard of it who haven’t heard of any other tool,” she says. Granger and Pérez do not know how many people have tried their software, but say that traffic to the website suggests that roughly 500,000–1.5 million people actively use the program. Nelson says that it is the best-designed of the digital notebooks, and attracts many users because it is free and open source. The application also benefits from the popularity of the Python language, which boasts a robust scientific community that meets for an annual international conference, and is (relatively) easy for novice programmers to learn.

➔ **NATURE.COM**  
For more on scientific software, apps and online tools, visit: [nature.com/toolbox](http://nature.com/toolbox)

Although a growing number of researchers are publishing their notebooks alongside papers

(see [go.nature.com/mqonbm](http://go.nature.com/mqonbm) for examples), it may still be some time before journals accept the documents as full journal articles.

A handful of IPython notebooks have been published as books, and many professors use the program to make interactive curricula. But so far, the notebooks seem to have been published only as addenda to papers — often to provide analysis code and additional explanation in method sections.

“Publishers, I would say, still aren’t convinced that they want to go the whole way,” says Granger. The data format may be too new, he says, for journals to recognize the notebook as an official document format, such as html or pdf. But the IPython team has begun talks with a few publications.

**START FROM SCRATCH**

Most IPython notebook users are skilled programmers, but experts are helping to introduce beginners to coding through the software (see ‘IPython for beginners’). Yan Song, a post-doc at the University of California, San Diego, had no programming experience until about three months ago. She works on the ‘wet’ side of a cellular and molecular medicine lab, where she designs experiments and collects data that computational scientists — on the ‘dry’ side — help her to mine for information.

Song looks for changes in RNA expression in mouse and human stem cells as they differentiate into various types of neuron. In the past, she used Excel to compare expression patterns between groups of cells at different developmental stages. But earlier this year, she began to examine RNA sequencing data from single cells and her data sets exploded in size and complexity. Instead of analysing a few groups of cells, she had to compare hundreds of cells at once, and in each one she examined around 1,500 separate genes related to neural development.

Olga Botvinnik, a bioinformatics graduate student in that lab, started to generate the results in an IPython notebook, so Song began playing with the analysis code — out of a mixture of curiosity and impatience, she says. “It seemed to be an easy interface. You can code one line and you can see whether it works right away.”

Within a few weeks, Song had picked up some basic IPython programming skills, finding support through online tutorials and messageboards. Botvinnik has also written some customized menus and widgets to let Song explore her data using different clustering algorithms.

Song still relies on Botvinnik for help with intensive computational analyses, but says that she is now starting to explore the data on her own, using her biological knowledge to examine particular subsets of cells or genes, which she can suggest to Botvinnik as leads for further analysis. “We used to speak two different languages. I would talk about the biology and she would talk about coding. Now we have common ground; we can communicate to each other better. This accelerates our research,” she says. ■



# CAREERS

**NATUREJOBS FACEBOOK** Links to and articles on job tips and info [www.facebook.com/naturejobs](http://www.facebook.com/naturejobs)

**NATUREJOBS BLOG** The latest on careers news and tips [blogs.nature.com/naturejobs](http://blogs.nature.com/naturejobs)

**NATUREJOBS** For the latest career listings and advice [www.naturejobs.com](http://www.naturejobs.com)



ILLUSTRATIONS BY LUCIANO LOZANO/GETTY

or urgency. These problems inevitably result from applicants' failure to allow themselves enough time to write the proposal and to circulate it to colleagues, advisers and department heads for feedback. This pattern is repeated twice a year, every year, when our submission deadlines approach.

I am also amazed anew each time to find that most of our grant-seekers wait until five minutes to midnight to meet our published deadlines. Yes, we circulate and read last-minute applications, but we have less time to ask for clarification or extra information in this flood tide because the clock is ticking for our next board meeting. And foundations always get more good proposals than they can fund.

## ALLOW ENOUGH TIME

Scientists cannot plan their protocols for hypotheses, goals, controls, methodologies and analyses and then write, edit, proofread, copy-edit, chart, graph and lay out their work effectively and error-free without input from colleagues. If your institution does not have an internal review process, then you are already at a disadvantage in the heated competition for funding and should take the initiative and ask your co-workers to critique your efforts. This means finishing your draft well in advance of submission dates. You need to give yourself enough time to polish your proposal — and to get useful, meaningful input on it. Two months ahead may not be too early.

Scientists are not trained as writers, and their applications would often benefit from editing. Although the proposals we receive do not usually contain vocabulary or grammatical errors, they are frequently repetitive. Often, the very point of the research is deeply buried in the proposal and does not emerge until well after a lengthy discussion of the background, when it should appear in a brief introduction or a summary at the top of the document. I also find with surprising frequency that important information — the current population of an endangered species, for example, or why a species should be studied at all — is missing, either because applicants think “everyone knows that”, or because details are lost when the focus is on the big picture. Do not make this mistake — it results from being too close to your own work to read it objectively, and you can avoid it by seeking comment and by scheduling enough time into the process to let the ►

## COLUMN

# It takes time and a team to win grants

*Start and finish early, seek feedback and file before deadline, says Ingrid Eisenstadter.*

**I**n the 25 years that I have been director of grants for a small family foundation that supports scientific research, I have reviewed a few thousand grant proposals. All our applicants are people who were bright enough to get PhDs and MDs, but the proposals we receive tend to share the same flaws, whether they come from recent

graduates or from researchers with years of experience.

Applicants often submit proposals in which the research protocol is insufficiently planned or explained. The language is sometimes too technical for reviewers who do not specialize in that discipline. The proposal text can be wordy or fails to convey the study's novelty

► proposal rest for a week or two and then rereading it with a fresh eye.

For example, we once received a request for funds to study an endangered primate. Yet the applicant did not mention until halfway through the proposal that fewer than 300 of these animals had not yet been wiped out by *Homo sapiens*. I called the applicant and suggested that they add that number to the proposal's title before I circulated it. The person said, "Oh! Right!", laughed, quickly resubmitted — and was funded. Had that single mention of this crucial number been missing altogether, there is no knowing what the result would have been, especially if it had been submitted during our biannual flood tide.

### DON'T FORGET THE DETAILS

Inadequately planned or poorly explained research are other common problems. These, too, can be corrected with input from a neighbouring bench or two. For instance, we recently heard from a scientist who put considerable effort into a proposal to study the effects of forest fragmentation on a naturally occurring hybrid tree. The regions to be compared included a swamp that had been drained to become a farm in colonial times; an artificially created urban park; a new suburban park; and others — all without any explanation for these site selections or, for that matter, of the broader significance of this choice of the tree for study. The application cited no references. It was not funded.

Another time, we received a proposal for a genetic study that clearly needed to include epigenetics, considering the speed with which the change under investigation had taken place.

These proposals, as written, should never have left their home institutions. Had they been reviewed by the applicants' colleagues, perhaps these basic problems would have been spotted in time for the applicants to recast their protocols.

Almost all the grant applications that we receive seek our maximum funding level, or an amount very close to it. In addition to coverage for supplies, lab fees, travel and other outlays, most applicants want salary support — whether or not they are obliged to raise their own salaries — and a contribution to the overheads of their institutions. I have spoken to many employees at other small- to medium-sized foundations who say the same thing. 'Shooting for the Moon' does not enhance your chances of funding.

Many foundations' websites provide a history of the grants that were funded, and you should use those figures to guide your budget decisions. In the United States, some foundations' websites include their annual federal tax forms (called Form 990), which list all the grants awarded each year. You should research this information well in advance of writing a grant, and tailor your application to the standards of the foundation to which you are applying. Include a budget justification that explains each expense, so that the foundation knows what it is paying for. There is no procedure that bars bargain-hunters from serving on foundation boards, and strong proposals that seek less rather than more may be more favourably viewed.

About 20 years ago, we got a request for US\$700 — which had never happened before and has not happened since — from a researcher who wanted fieldwork support to study the threatened blue copper butterfly (*Lycaena heteronea*), a beauty of the western United States. We awarded the grant without hesitation. A year later, our lepidopterist reported that his butterfly was not as threatened as he feared. Our board of directors was as delighted as he was.

It is a mistake to assume that all the grant reviewers at non-government funding organizations who will ultimately vote on your proposal are scientists. Unless you are writing about particle entanglement, use plain and non-technical language whenever possible. If you do not, or if your topic makes that impossible, your proposal may well go to referees, leaving grant decision-makers to depend on someone else's opinion. If you start the application process with a letter of inquiry — a brief memo

that discusses your work — you should already have planned your research and should know how much you need in funding. Consider including the amount of your grant request in the letter, because if you later submit a full proposal that asks for much more than what the board had expected, your chances of funding are diminished.

Funders do understand that letters of inquiry are sometimes vague about research plans because investigators are seeking expressions of interest before taking the time to prepare a detailed protocol and full proposal. Nonetheless, early-career researchers as well as senior scientists should realize that it is difficult for vague letters of inquiry

to compete with those that make it clear that there is a complete research plan behind them.

Subheads are an important navigation tool for proposal evaluators. Use them to highlight the importance and novelty of your work, and be clear; for example, write 'This species is now endangered', rather than 'This species is now on the IUCN Red List'. Even when you are following a specific question-

**"Unless you're writing about particle entanglement, use plain and non-technical language whenever possible."**

and-answer formula created by the grant-giver, consider adding subheads to emphasize your proposal's strengths and urgency. Some grant-givers have such a strict set of questions that there is little opportunity

to explain the goal or necessity of your work. If so, add an 'introduction' subhead to bring out these points, and if you keep the accompanying text to a few sentences that enable you to address the issue missing from the one-size-fits-all questionnaire, you may not get into trouble.

### ILLUSTRATE WELL

Similarly, photos, charts and graphs should highlight and emphasize the importance and significance of your work. Now that technology has facilitated the use of photos in grant proposals, we are seeing them more often. If you plan to use them, remember that they should be informational, not decorative. You also need to remember that evaluators will look at photos, charts, graphs and their captions before they read the text on that page, so captions should underscore the significance of the work.

It is also important to explain the future ramifications of your research after you complete the current phase for which you are seeking funding. That information is often missing. If your research will facilitate others' investigations, or will continue in some other way to ripple in the water, then say so, whether your proposed research programme is basic or applied. Do not leave the evaluators of your proposal to have to figure this out.

Most of the researchers and institutions that we have funded end all communication with us when we get their final reports. But every now and again, wise researchers send us copies of their publications as the years pass, along with a note that explains the relevance of the studies to the earlier work that we funded. This practice boosts your chances of success should you ever want to seek funding again. It is also the courteous thing to do. ■

**Ingrid Eisenstadter** is director of grants for The Eppley Foundation for Research in New York.



CLAIRE WELSH/NATURE



# BOUNDARY WATERS

*Dive in.*

BY MARISSA LINGEN

She had not expected the howls.

All babies cried, she knew — or at least she thought she knew. Somehow it was different when it was her own, and all her theories about calmly going about her business went out the window. But the modifications for gills were supposed to make time in the water a joy, a delight — she had pictured a cooing, fascinated baby, not this shrieking red monster.

She had pictured almost a modern selkie baby, large-eyed and at peace in the waves, wise beyond her days, able to cope with the modern world of rising tides and temperatures.

And that baby did come, eventually — after 20 minutes or so in the waves, settling into breathing in the water and then wailing like the torments of the damned *again* when they took her out.

"I wish she'd make up her mind," she said, joggling the baby in the eternal parental soothing dance, still half-towelled and tangle-haired herself.

"Maybe we should just ... leave the swimming alone for awhile," he said. "If she doesn't like it."

They took the baby to the doctor, and also to the genetic-modification counselor, to see if something was wrong with the amphibious changes they'd selected. Corvallis was close enough to the coast — even before the sea level had risen, there had been professors with beach homes — that there were others with the same modification package. The clinics were experienced.

They found no problems. They shrugged and told the parents not to let her skin get too dry, not to use harsh perfumed lotions on the gill region.

"Harsh perfumed lotions on a baby," she snorted in the car.

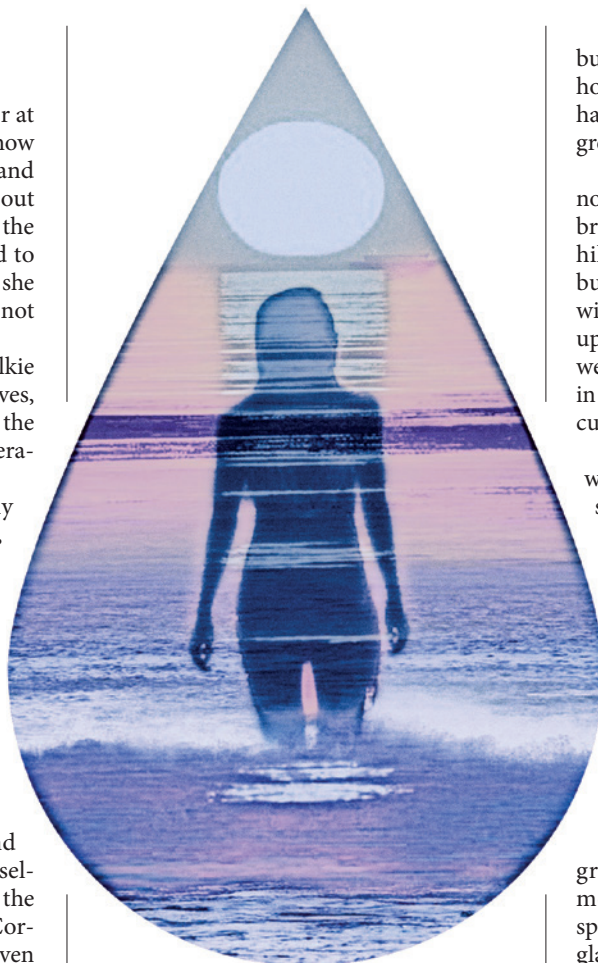
"Some people don't know," he soothed.

"She needs to ... she needs to be comfortable with her heritage."

"I'll teach her to cook nopales so they don't get slimy," he promised.

This was not what she meant.

The requisite long bath times were a struggle at first. Then it was eerie and upsetting to find an eight-year-old lying perfectly still under the water for hours at a time. When there was a waterproof tablet in the bath, that at least



seemed like something. When it was just the child and the water, perfectly still, the laughing selkie baby of imagination seemed never more distant.

"Kids," said her mother. "You shouldn't think you know what you're getting, because you never, ever do. You were just like that, hours at a time."

"In the *tub*?" she said. "I never."

"No, up in that tree, the big maple on the side yard."

"That was different."

Her mother snorted.

She was 22, and it was a long drive inland, parents and U-Haul and all. She had thought she would just go on her own, but her parents said there were no easy flights and the bus was a mess and ... there were all sorts of reasons to take her, and once they were taking her, why, may as well take the old second-hand kitchen table they'd outgrown, some chairs, all the things they imagined she'd need.

They'd never expected she'd go so far east, but once she'd tasted fresh water, she was home, and there was no going back to the harsh tang of salt, the familiar and foreign greeting of sea stars and kelp.

They helped her unpack, shelving her limnology textbooks in the duplex she'd found; brick and white clapboard up in the steep hills, identical to all the other blue-collar buildings perched there, shocking them with its cheap rent. They'd helped her stock up on all the food she'd need to keep her weight up, keep her fat stores up to survive in the water, even with the wet suits specially cut around her gills.

They wanted to tour the campus. They wanted to see the piers, the ore docks, the sand and black-stone beaches. They tried to talk about how the forests were like the forests back in Oregon, but they were not like. Nothing smelled like. The water, when they were gone, would not taste like.

When they drove off, she sat and cried, and her tears tasted treacherously, so unfairly, of salt. No matter how long she spent under the surface, spring through autumn, that wouldn't change.

There would be others like her, in her graduate programme, others who were modified to be here, quiet and loath to speak, slow moving, large, glad of the cold, glad of the silences. They found pockets of zebra mussels. They fled the coasts. They analysed traffic.

Their parents had no idea what to do with them.

She walked down to the beach. Most of the people on it were unmodified, but there were enough like her that they were too polite to watch, used to it. She slipped into the waves, and they muffled her cries.

Even in her own kind of water, she still cried out. She had never stopped crying out, with the change. But the freshness helped, sliding into the depth helped. The quiet helped.

She crouched in the cool, cool corner and ate her fish, thinking about the shipping lanes and what to do next.

She had never been able to explain to her mother that all the modification could do was make the transitions possible. It could never make them easy. ■

**Marissa Lingen** has published more than 100 short stories in venues such as *Analog*, *Lightspeed* and *Tor.com*.

JACEY

➔ NATURE.COM

Follow Futures:

Twitter @NatureFutures

Facebook go.nature.com/mtoodm